

THE SCIENTIFIC
STUDY
OF
EDUCATIONAL
PROBLEMS

MONROE
AND
ENGELHART



Division LB1028

Section .M 75

Experimental Education Series

EDITED BY M. V. AND H. E. O'SHEA

**THE SCIENTIFIC STUDY OF
EDUCATIONAL PROBLEMS**

EXPERIMENTAL EDUCATION SERIES

EDITED BY M. V. AND H. E. O'SHEA

CHILD PSYCHOLOGY.

By GEORGE D. STODDARD, Ph.D., Professor of Psychology and Director, and BETH L. WELLMAN, Ph.D., Associate Professor of Psychology, Iowa Child Welfare Research Station, University of Iowa.

A MANUAL OF CHILD PSYCHOLOGY.

By the same authors.

THE DIAGNOSIS AND TREATMENT OF BEHAVIOR-PROBLEM CHILDREN.

By HARRY J. BAKER, Ph.D., Director, and VIRGINIA TRAPHAGEN, M.A., Mental Examiner, Psychological Clinic, Detroit Public Schools.

ENRICHING THE CURRICULUM FOR GIFTED CHILDREN.

By W. J. OSBURN, Professor of School Administration, The State University of Ohio, and Director of Educational Research, Ohio State Department of Education, and BEN J. ROHAN, Superintendent of Schools, Appleton, Wisconsin.

FITTING THE SCHOOL TO THE CHILD.

By ELISABETH IRWIN, Psychologist, Public Education Association of New York City, and LOUIS A. MARKS, Member Board of Examiners, Board of Education, New York City.

THE FUNDAMENTALS OF STATISTICS.

By L. L. THURSTONE, Ph.D., Bureau of Public Personnel Administration, Washington, D. C.

GIFTED CHILDREN: THEIR NATURE AND NURTURE.

By LETA S. HOLLINGWORTH, Ph.D., Associate Professor of Education, Teachers College, Columbia University.

HOW TO EXPERIMENT IN EDUCATION.

By WILLIAM A. MCCALL, Ph.D., Associate Professor of Education, Teachers College, Columbia University.

MODERN PSYCHOLOGIES AND EDUCATION.

By CLARENCE E. RAGSDALE, Ph.D., Assistant Professor of Education, The University of Wisconsin.

PRINCIPLES OF MUSICAL EDUCATION.

By JAMES L. MURSELL, Ph.D., Department of Education, Lawrence College.

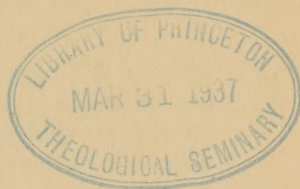
SPECIAL TALENTS AND DEFECTS.

By LETA S. HOLLINGWORTH, Ph.D., Associate Professor of Education, Teachers College, Columbia University.

THE SCIENTIFIC STUDY OF EDUCATIONAL PROBLEMS.

By WALTER S. MONROE, Professor of Education and Director, Bureau of Educational Research, University of Illinois, and MAX D. ENGELHART, Director, Department of Examinations, Chicago City Junior Colleges.

THE SCIENTIFIC STUDY OF EDUCATIONAL PROBLEMS



BY

WALTER S. MONROE

PROFESSOR OF EDUCATION AND DIRECTOR, BUREAU OF
EDUCATIONAL RESEARCH, UNIVERSITY OF ILLINOIS

AND

MAX D. ENGELHART

DIRECTOR, DEPARTMENT OF EXAMINATIONS,
CHICAGO CITY JUNIOR COLLEGES

NEW YORK
THE MACMILLAN COMPANY
1936

COPYRIGHT, 1936,
By THE MACMILLAN COMPANY

ALL RIGHTS RESERVED—NO PART OF THIS BOOK MAY BE
REPRODUCED IN ANY FORM WITHOUT PERMISSION IN WRITING
FROM THE PUBLISHER, EXCEPT BY A REVIEWER WHO WISHES
TO QUOTE BRIEF PASSAGES IN CONNECTION WITH A REVIEW
WRITTEN FOR INCLUSION IN MAGAZINE OR NEWSPAPER

Published December, 1936

SET UP AND ELECTROTYPED BY T. MOREY & SON

: Printed in the United States of America :

PREFACE

This volume is addressed to a varied audience. It is intended to provide a basic text for graduate students in education and others who are interested in learning how to study educational problems scientifically. It is intended to serve also as a source of information for research workers. A third group consists of the consumers of educational research. These include superintendents, principals, supervisors, teachers, and others who endeavor to ascertain what research has accomplished in the field of education and to interpret the findings with respect to theory or practice. This large consumer group is addressed for the purpose of engendering attitudes and other abilities necessary for reading intelligently reports of studies in educational periodicals and other publications. Recognition of these three groups has created problems in determining the content which would not have arisen if the volume had been addressed to a more homogeneous audience. The degree of wisdom that the authors have exercised in the selection of topics and in the treatment of them can be determined only by those who make use of the results of their efforts.

It seemed wise to include a fairly comprehensive treatment of the statistical techniques employed in educational research. For those most likely to be used by graduate students and other persons interested in attempting studies in the field of education, the treatment is sufficiently detailed so that it should not be necessary to consult other sources. For the less commonly employed techniques, the treatment is somewhat abbreviated, but references are given to sources from which additional information may be secured. Considerable space is given to the interpretation of statistics, a topic that has not received adequate attention in texts on educational statistics. It is hoped that the volume may find a place as a text for courses

in statistics as well as courses devoted to the study of educational research.

The volume is critical. Educational research is a new field and, as is to be expected, its techniques have been only imperfectly understood by many of those who have employed them. In our enthusiasm in attempting to deal with educational problems scientifically, many errors have been made. By calling attention to the limitations of educational data, and by revealing difficulties encountered in educational research, the authors hope they have succeeded in presenting an adequate picture of the situation and thereby have provided a basis for more constructive research endeavors. The authors have faith in the possibilities of educational research and they hope the reader will develop an intelligent optimism.

In view of the number of texts on educational statistics and the attempts that have been made to describe educational research, a new venture in this field should justify itself as a contribution. The evaluation of the present volume will be made by those who examine its pages systematically. But the authors may be permitted the privilege of mentioning certain features that they believe make it distinctive. The interpretation of the findings of research in the light of their dependability has been emphasized by directing attention to the limitations of educational data and of the techniques employed in handling them. Descriptions of statistical techniques may be found in other writings but the competent reader who examines the volume carefully will note a number of details, especially in Chapters X and XI that are not generally known. Another feature is the comprehensiveness of the volume which is indicated by the topical index. The limitations of space prohibit more than a brief mention of a number of the topics listed, but in most such cases references are given so that the interested reader may consult other sources. The organization of the volume, together with this index, should make it an effective source of information. Finally, the authors have attempted, especially in the concluding chapter, to indicate the types of

studies to which workers must direct their efforts if a science of education is to be built up.

It will be obvious to the reader that the authors are indebted to many research workers and other writers in this field. They have attempted to acknowledge this indebtedness by numerous references to sources, but these citations do not adequately indicate the extent to which their thinking has been stimulated and guided by other persons. The limitations of space prohibit listing the names of this large group, but the senior author desires to make a general acknowledgment of his indebtedness to his students, especially those on the graduate level, and to the members of the staff of the Bureau of Educational Research during the period, 1921 to 1931. Specific mention should be made of Dr. P. T. Orata and Dr. D. B. Stuit, both of whom read portions of the manuscript and of Dr. Harold Gulliksen and Mr. L. R. Tucker who were consulted in regard to factor analysis. Several findings from the dissertation of Dr. Stuit are reproduced in Chapter XI. Finally, the authors are grateful to Miss Neva M. Covey for her intelligent typing and checking of the manuscript.

WALTER S. MONROE
MAX D. ENGELHART

September, 1936

TABLE OF CONTENTS

CHAPTER	PAGE
I. INTRODUCTION	1
II. EDUCATIONAL PROBLEMS AND THEIR DEFINITION	13
III. COLLECTING THE DATA SPECIFIED BY A PROBLEM—BASIC TECHNIQUES	28
IV. ELEMENTARY TECHNIQUES FOR HANDLING DATA	61
A. Scales and Calculations	61
B. Statistics of Frequency Distributions	65
C. Calculation of Comparable Measures	82
D. Statistics of Relationship	85
E. Measures of the Effect of Chance in Random Sampling	103
F. Interpretation of Statistics	112
V. THE FAULTS OF DATA AND THEIR EFFECTS	121
A. Types of Data Faults	122
B. Errors of Measurement	130
C. Errors of Validity	144
D. Effect of Data Faults	149
VI. STUDYING THE PAST IN EDUCATION	159
VII. CONSTRUCTING MEASURING INSTRUMENTS	171
A. General Principles	171
B. Details of Test Construction	178
C. Description of Test Performances	188
D. Estimating the Excellence of Measuring Instruments	198
E. Improvement of Tests	208
VIII. STUDYING CURRENT CONDITIONS OR PRACTICES	214
A. Techniques of Surveys	216
B. Reporting and Interpreting Findings	242

CHAPTER	PAGE
IX. STUDYING THE EFFECT OF A SPECIFIED CHANGE IN A GIVEN CAUSE	270
A. Procedure of Experimentation	274
B. Interpretation of Experimental Findings.	306
X. STUDYING PROBLEMS OF PREDICTION	323
A. Methods of Making Predictions	323
B. Measures of the Accuracy of Predictions.	333
C. Efficiency of Predictions in Practice	350
XI. IDENTIFYING AND STUDYING CAUSE AND EFFECT RELATIONSHIPS	366
A. The Nature of Relationship	366
B. Measurement of the Contributions	371
C. Partial Correlation	377
D. Regression Equations and Factor Analysis	389
XII. DETERMINING WHAT SHOULD BE	411
XIII. EVALUATING AND SYNTHESIZING EDUCA- TIONAL RESEARCH	436
XIV. PROGRESS TOWARD A SCIENCE OF EDUCATION	453
APPENDIX—STATISTICAL SYMBOLS	477
AUTHOR INDEX	487
TOPIC INDEX	499

LIST OF FIGURES

FIGURE	PAGE
1. Calculation of a Pearson Product-Moment Coefficient of Correlation	91
2. A Tabulation Sheet for Four Items of Information, Sex, High School Class, Age, and IQ	230
3. Bar Graph of Variability of English Achievement in Several Colleges	242
4. Curves for Predicting New Stanford Reading Test Scores from Scores on the Otis Self-Administering Test of Mental Ability	328
5. Different Measures of Efficiency of Prediction for Values of r_{01}	347
6. Path Coefficient Diagram for Three Independent Variables .	396

LIST OF TABLES

TABLE	PAGE
I. Showing the Tabulation of a Frequency Distribution . . .	68
II. Illustrating the Calculation of the Mean from a Frequency Distribution	70
III. Illustrating the Calculation of the Standard Deviation of a Frequency Distribution	76
IV. Illustrating a Correlation Table	87
V. The Probabilities that the Value of a Statistic for a Universe Lies within the Interval Formed by Subtracting and Adding a Multiple of Its Probable Error	106
VI. Distribution of Differences between Scores Yielded by Two Applications of Monroe's General Survey Scale in Arithmetic to a Group of Fifth-Grade Pupils	126
VII. Two Sets of Gains in Achievement Which Indicate the Presence of Constant Errors in Certain Sets of Scores, Fifth Grade	138
VIII. Subjectivity of Scoring Reproductions by the Word-Counting Method	141
IX. Values of Coefficient of Reliability r_{1I} and Corresponding Values of Probable Error of Measurement $.6745\sigma_1\sqrt{1-r_{1I}}$	206
X. Average Spelling Achievement and Minutes per Day Devoted to Spelling, Seventh Grade. After Rice	271
XI. Showing for Various Values of r_{01} , the Per Cent of Predictions Whose Error Is Not Greater than the Amount Indicated.	339
XII. Efficiency of Prediction of X_0 by Means of Regression Equation, for Various Values of r_{01} or $R_{0.12\dots n}$	343
XIII. Efficiency of Prediction of x_0 , Where $\frac{\sigma_0}{\sigma_1} x_1$ Is Taken as Evidence of x_0 , for Various Values of r_{01}	344

TABLE	PAGE
XIV. Efficiency of Prediction of \bar{X}_∞ by Means of Regression Equations, for Various Values of r_{01} or $R_{0.12 \dots n}$ and r_{00}	346
XV. Illustrative Values of Two Correlated Variables Showing the Common Factor	368
XVI. Number of Doctors' Degrees in Education	461

EDITOR'S INTRODUCTION

It is a widely recognized truth that the characteristics of adult society are dependent more largely upon the education given to its children and to its youths than upon any other single factor. Consequently, nothing, certainly, can be of greater concern to society than genuinely sound educational thinking. The methods of thinking which have revealed what is known about the structure of the universe through astronomy, physics, chemistry, and biology, have in recent years been applied to the problems of education with ever increasing fruitfulness.

Professor Monroe and Dr. Engelhart in *The Scientific Study of Educational Problems* have put these fruitful procedures into readily understandable terms in order that they may be learned and utilized by the greatest possible number of educational workers. The authors have endeavored to cast the material in the form in which educators approach their own subject matter, rather than to arrange it in some abstract order which might conceal scientific processes in educational thinking from all but the most highly trained technicians.

It is the particular hope of the authors and of the editor that not only will those persons who go forward to discover new truths in education be stimulated and assisted by this volume, but also that the group of educators ultimately most important, the administrators and the teachers who actually deal with children and with young people, will receive substantial help from this volume in understanding and interpreting the investigations of research workers. Scientific data in education would have little, if any, significance if they were to remain only on shelves of research monographs; they become of importance to society when they are finally utilized by a wise teacher to improve the experiences and enrich the opportunities of her pupils.

HARRIET E. O'SHEA

PURDUE UNIVERSITY
September, 1936

THE SCIENTIFIC STUDY OF EDUCATIONAL PROBLEMS

CHAPTER I

INTRODUCTION

The meaning of educational research. A person confronted with a question may accept the first answer that occurs to him; he may derive an answer from previous experience or casual observation; he may consult other persons in regard to the answer; or he may seek his answer to the question in the published opinions of others. Such means of answering educational questions are not those of educational research. In contrast educational research may be thought of as the total procedure employed in collecting, handling, and interpreting data for the purpose of arriving at dependable answers to questions about education.¹

Sometimes the question for which an answer is sought is so simple that the total activity of answering it involves little more than collecting the required data and applying simple arithmetical techniques. For example, suppose a city superintendent wishes to ascertain the grade enrollments of his system. He asks his teachers to report the number of pupils enrolled with a designation of their grade classification. The numbers reported are summarized and the totals constitute the answer to the question. If such activity is labeled "educational research," it is obvious that the term will have a very general meaning.

¹ This statement requires modification when educational research is regarded as inclusive of the philosophical type of inquiry. The conclusions of philosophical educational research are in the nature of decisions with respect to "what should be." For further discussion, see Chapter XII.

Examination of educational writings reveals that although the term is not applied in such cases, the designation of "educational research," is frequently given to relatively simple and routine investigations such as surveys of current conditions or practices. This custom is unfortunate. The usage of the term should be limited so that research in education will be comparable in its essential characteristics with research in other fields.

Illustrations of educational research. Before attempting to point out the essential characteristics of educational research, a few studies will be described briefly. If feasible, the reader should consult the references cited and study the complete reports for a more adequate understanding of the nature of these researches.

1. *The Department of Superintendence study of the status of the superintendent.*¹ The problems of this survey investigation were as follows: "(1) To determine the status of the superintendent of schools with reference to training, experience, and tenure; (2) To determine the facts regarding the financial compensation of the superintendent of schools; (3) To determine the professional activities in which the superintendent of schools is engaged; (4) To determine as far as possible the economic status of the superintendent; (5) To determine the interrelationships between elements mentioned above."²

Data were collected chiefly by means of a questionnaire mailed to city superintendents under the auspices of a committee of the Department of Superintendence of the National Education Association. The questionnaire is characterized by questions asking for facts rather than expressions of opinion. Returns were received from 1181 superintendents—approximately 40 per cent of the superintendents in cities over 2500 in population in the United States. Some additional data were collected from the Educational Directory of the United States Bureau

¹ Chadsey, C. E., et al. "The Status of the Superintendent," *First Yearbook of the Department of Superintendence*. Washington: National Education Association, 1923. 206 pp.

² *Ibid.*, p. 10.

of Education. Although the statistical techniques employed in summarizing the data are very simple, the number of questionnaire returns was so large that mechanical means of tabulation were employed.

2. *The Judd and Buswell study of the nature of eye-movements in silent reading.*¹ The problem of this laboratory study was to determine the characteristics of the eye-movements made in certain types of reading. The data were collected by securing a photographic record of the eye-movements of a number of subjects while engaging in reading silently various types of material and in reading the same material in response to various types of requests.² From the photographic record on a moving picture film the investigators were able to determine the following facts for each line of the text read: (1) number of eye-movements, (2) their direction (forward or regressive), and (3) the length of each period of fixation.

In the Judd and Buswell study photographic records were secured for various types of silent reading—rapid, superficial reading; slow, careful reading preparatory to answering questions; reading when difficult words are encountered; reading an easy poem; reading a passage to be reproduced; reading with grammatical analysis; reading a foreign language. The numerical data derived from the photographic records were assembled to show the variations in the performance of a subject when engaging in the different types of reading and to show the performances of different subjects for a given type of reading.

It should be noted that the data collected are highly objective,

¹ Judd, C. H., and Buswell, G. T. "Silent Reading: A Study of the Various Types," *Supplementary Educational Monographs*, No. 23. Chicago: University of Chicago, 1922. 160 pp.

² The apparatus used for securing the photographic record is somewhat complicated and need not be described here. In reading the movement of the eyes is not a smooth rotation from the left to the right but a series of short rapid movements and brief pauses or *periods of fixation*. Furthermore, there may be some movements from right to left which are known as *regressive*.

For a description of the apparatus see:

Gray, C. T. "Types of Reading Ability as Exhibited through Tests and Laboratory Experiments," *Supplementary Educational Monographs*, Vol. 1, No. 5. Chicago: University of Chicago, 1917, pp. 83-91.

that is, there was very little opportunity for them to be influenced by any prejudices or desires of the investigators. Another person using the same subjects and following the same procedure would have secured approximately the same data. The data are also highly accurate and valid. It may be noted also that only simple arithmetical techniques were employed in handling the data.

3. *The Newark phonics experiment.*¹ The problem of this controlled experiment was to determine the relative effectiveness of teaching beginning reading with instruction in phonics and without instruction in phonics. The characteristic of the study is the employment of the experimental technique in collecting the data. One group of pupils was given instruction in phonics and another group with nearly identical status with respect to reading ability was instructed without training in phonics. At the end of five months four reading tests were administered to all pupils. The pupils were tested again as they completed the work of grade IA and again as they completed IIB. The scores obtained by administering these tests constitute the data of the experiment. Collecting the data, however, involved more than merely administering tests. Setting up the experiment so that the experimental and control groups would be equivalent in reading ability and conducting the study so that the other non-experimental factors would be adequately controlled are essential phases of the process. Differences in average gains in reading ability of the experimental group and the control group were computed. The statistical procedures, however, are relatively simple.

4. *A prediction study.*² Douglass has reported a study in which he sought to ascertain the value of high school record, intelligence test score, and certain other data for predicting the success

¹ Sexton, E. K., and Herron, J. S. "The Newark Phonics Experiment," *Elementary School Journal*, 28: 690-701, May, 1928.

² Douglass, H. R. "The Relation of High School Preparation and Certain Other Factors to Academic Success at the University of Oregon," *University of Oregon Publication*, Vol. 3, No. 1. Eugene, Oregon: University of Oregon Press, 1931. 61 pp.

of students entering college and to derive a formula for using the data shown to be valuable. Most of the data were copied from records on file in the registrar's office at the University of Oregon. Certain information was obtained from the state educational directory. An intelligence test was administered to secure a measure of the intelligence of the entrants. In studying the data both partial correlation and multiple correlation were employed. Hence, the statistical procedures in this study were more intricate than those employed in the three preceding ones.

5. *The Heilman study of the relative influence of certain hereditary and environmental factors on educational achievement.*¹ The problem of this investigation was that of determining "the relative influence upon scholastic achievement of mental age, school attendance, and socio-economic status of the home."² The problem has also been stated by Heilman as that of determining "which of three factors, mental age, school attendance, or the socio-economic status of the home, had the greatest weight in producing individual differences in the educational age of ten-year-old children."³

The following data were collected for 828 ten-year-old children: (1) Educational age in months; (2) Mental age in months; (3) Life age in months; (4) School attendance in days for each grade (and for the kindergarten); (5) A measure of the socio-economic status of the home; (6) Date of entering the first grade (and of the kindergarten for those children who had kindergarten training). Educational ages were computed from scores on the Stanford Achievement Test. Mental ages of the children were secured by employing the Stanford Revision of the Binet Intelligence Scale. Age and attendance data for the children were obtained from the records of the

¹ Heilman, J. D. "The Relative Influence upon Educational Achievement of Some Hereditary and Environmental Factors," *Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1928, pp. 35-65.

² *Ibid.*, p. 35.

³ Heilman, J. D. "Factors Determining Achievement and Grade Location," *Pedagogical Seminary and Journal of Genetic Psychology*, 36: 435, September, 1929.

Denver schools. Data pertaining to the socio-economic status of the homes of the children were obtained by means of a revision of the Chapman-Sims Socio-economic Scale.

In handling his data Heilman found it necessary to postulate the directions of causation. On *a priori* grounds he inferred that educational age is directly influenced by school attendance, mental age, and socio-economic status. He assumed also that mental age indirectly influences educational age through its influence on socio-economic status and on school attendance; that socio-economic status indirectly influences educational age through school attendance; and that no significant reciprocal influences operated. In the statistical treatment of his data Heilman computed coefficients of correlation between educational age and mental age, educational age and socio-economic status, educational age and chronological age, and so on for each of the possible combinations of the paired sets of measures. He then applied the technique of partial correlation as a means of eliminating chronological age as a factor. The method of "path coefficients" devised by Wright¹ was introduced as a means of determining a measure of the relative contributions of the remaining factors to individual differences in the educational age of ten-year-old school children.

6. *The Hockett study to determine the more significant political, economic, and social problems of American life.*² This study by Hockett was made for the purpose of determining what political, social, and economic problems and issues of American life about which children should become able to think intelligently. As a basis for answering this question, which asks what should be, he assumed that in a commonwealth such as ours the citizens need to understand the society in which they live and to be able to deal with the problems which they are likely to face.

¹ Wright, Sewall. "Correlation and Causation," *Journal of Agricultural Research*, 20: 557-85, January, 1921.

² Hockett, J. A. "A Determination of the Major Social Problems of American Life," *Teachers College, Columbia University Contributions to Education*, No. 281. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 101 pp.

He further postulated that a consensus of expert opinion would afford the best answer to this question. His first step, therefore, was to determine the persons who might properly be recognized as "frontier thinkers" and their writings which represent "in the highest degree penetrating insight and critical analysis of contemporary life and problems." In this step of his work Hockett sought the help of "150 specialists in the field of government, economics, sociology, law, the press, international affairs, immigration, geology, anthropology, and the field of artistic expression." With the aid of the information contributed by these specialists he selected twenty-two books for analysis.

The analysis of these twenty-two volumes was, of course, subjective, but Hockett's procedure was highly systematic and it is probable that the list of 396 issues which he identified constitutes a satisfactory determination of the consensus of opinion of the authors of the books analyzed. Hockett sought to validate the results of the analysis by a study of the reported events of American life over a period of five years. In identifying these events he analyzed the weekly summary of events under the heading of "current events" in the *Literary Digest* and the editorial comment on events appearing in the *Outlook*, the *Independent*, the *New Republic*, and the *Nation*.

Characteristics of educational research revealed in these illustrations. These illustrations of educational research reveal certain significant characteristics. In each case there is a problem expressible in the form of one or more questions and this problem when adequately defined served as a guide in collecting the needed data and in interpreting them. Although the nature of the data and the techniques employed in collecting them varied, the procedure may be described as systematic. Careful examination of the reports reveals that in each case there was a distinct effort to secure as accurate data as possible. With the exception of Hockett's study the method of collecting data was such that there was little or no opportunity for them to be influenced by any prejudice or preconceived opinion of

the investigator. Hockett's method of collecting data was subjective, but he shows that another analyst using his technique on a portion of the source material obtained very similar results. Although objective techniques for collecting data are desirable they are not essential. The essential requirement is that the process of collecting be systematic and that the resulting data be as accurate and as valid as possible.

Another characteristic, which is only suggested by the brief descriptions, is the attention to the faults of the data in interpreting them. Careful reading of the reports, however, makes clear that the several investigators recognized the faults of their data and endeavored to interpret the findings in accordance with these faults. This characteristic is perhaps most apparent in the Newark phonics experiment.

In each of these studies the investigator dealt with a limited collection of data, but he was interested in a problem more general than that indicated by his data. Judd and Buswell were not interested in the eye-movements of the particular subjects studied except as they were indicative of the eye-movements of readers in general. Sexton and Herron were interested in the findings of the Newark phonics experiment as a basis for advising teachers in regard to the instructional procedures in teaching reading in the primary grades. Heilman sought measures of the relative influence of certain factors upon the achievement of a particular group of pupils as a means of generalizing concerning the factors contributing to school achievement. Hence, we may note as a characteristic of educational research the interpretation of the findings as indicative of general conditions and relationships. Not infrequently the generalization is limited to a relatively restricted population or otherwise qualified and sometimes it is pointed out that generalization is perhaps not justified, but an investigation whose findings are applicable only to the particular population studied does not typify educational research. In fact, it may be maintained that such inquiries do not meet the requirements of educational research.

The characteristics of educational research may be epitomized as follows:

1. An investigation designed to afford a basis for generalizing with reference to educational theory or practice.
2. A problem whose definition serves as a guide in collecting the needed data.
3. Wisely planned and systematic collecting of data.
4. Critical interpretation of data with attention to their faults and any limitations implied by underlying assumptions.

The first of these characteristics is somewhat indefinite and hence difficult to apply, but it is suggested as a means of distinguishing between investigations that justify the designation of research and those that may more appropriately be called service studies or applications of research techniques in administration and teaching. Logically the definition of the problem is the first step in research. Frequently, the definition of the problem is modified and extended as the work progresses. Usually, however, a well defined problem may be regarded as a characteristic of educational research.

Sometimes we encounter the impression that educational research is largely a routine procedure and hence that when appropriate techniques are employed in collecting data the statistical treatment of them will yield the answers to the questions being studied. It is true that frequently there is much routine in connection with these two phases of educational research, but the total procedure cannot be described in terms of any definite techniques. It is essential that the collecting of the data be wisely planned with reference to the problem and that the process be systematic. With reference to the statistical treatment of the data, no criterion can be cited except the general one that the techniques applied be appropriate to the problem and to the data. In the studies described in the preceding pages, the techniques employed vary. In the Hockett study and in the Judd-Buswell study only the simplest of statistical techniques were employed. On the other hand, the Heilman study of the influence of certain hereditary and en-

vironmental factors on educational achievement called for the relatively elaborate statistical treatment including the application of the method of path coefficients.

If the designation of "poor research" is admitted, the fourth characteristic is possibly more of a criterion of the quality of the work than of research, but it refers to an essential phase of the total procedure of "good research." It is essential that a research worker be cognisant of the faults of his data and that he make due allowance for these faults in interpreting these findings. It is also essential that he be cognisant of the assumptions that may be implied by the problem or by the technique employed.¹

The plan of the following chapters. Chapter II deals with research problems and their definition in keeping with the postulate that "all educational research begins with a problem to be solved." Chapters III, IV, and V deal with standard methods of collecting data, standard methods of handling data, and the faults of data and their effects. The student should acquire from his study of Chapters II, III, IV, and V a knowledge of the general procedures of educational research from the selection and definition of a problem through the collection and handling of data to their interpretation in formulating conclusions and generalizations.

Chapters VI to XII differ from the preceding chapters in that each is devoted to a general type of research problem. While one purpose of these chapters is to provide information with

¹ These include fundamental assumptions such as that children learn or that human traits are sufficiently stable to make measurement of them meaningful. Assumptions of this type are essentially postulates and are to be distinguished from those introduced when a particular procedure is employed. The calculation of a mean from a frequency distribution introduces the assumption that the mid-point of an interval is the average of the measures within it. A comparison of the mean score of a class with other means or with announced norms introduces the assumption of comparability. The use of a probable error formula in general assumes the sample to which it is applied is a random one. The interested reader may consult the Index for references to the treatment of assumptions in later chapters. He should also read Gray, J. S. "Scientific Postulates for Educational Research," *Educational Administration and Supervision*, 19: 18-24, January, 1933, and Scates, D. E. "Types of Assumptions in Educational Research," *Journal of Educational Psychology*, 26: 350-66, May, 1935.

respect to the techniques appropriate for the several types of problems, each of these chapters has the additional purpose of presenting an indication of what has been accomplished by educational research in the field being considered.

Chapter XIII presents information with respect to the evaluation and summarization of educational research.

In Chapter XIV consideration is given to the progress that research workers have made toward a science of education and an attempt is made to point out the general limitations of current research and to indicate the lines along which efforts should be directed if a science of education is to be developed.

A list of statistical symbols is given as an appendix. The reader will find it helpful to consult this list when he encounters difficulty in understanding the symbolism in the following chapters.

Suggestions for using the volume as a text. The volume is somewhat encyclopedic in character so that it will be useful as a source of information. An instructor using it as a text will probably find that the objectives of the course suggest the omission of certain sections. For example, if the course is designed primarily to engender an elementary acquaintance with statistical procedure and ability to interpret the resulting statistics, most of the matter in Chapters II, III, VI, XII, XIII, and XIV would probably be omitted. If the purpose of the course is to engender an understanding of educational research and an ability to read reports of statistical studies intelligently, the omission of the more technical topics may be advisable, especially those in Chapters X and XI. If the students have previously had a course in statistical methods, Chapter IV may be omitted. If educational measurements is dealt with in a separate course, Chapter VII may be omitted.

SELECTED BIBLIOGRAPHY

ABELSON, H. H. *The Art of Educational Research: Its Problems and Procedures*. Yonkers-on-Hudson, New York: World Book Company, 1933. 332 pp.

12 STUDY OF EDUCATIONAL PROBLEMS

- ALEXANDER, CARTER. *School Statistics and Publicity*. Boston: Silver, Burdett and Company, 1919. 332 pp.
- ALMACK, J. C. *Research and Thesis Writing*. New York: Houghton Mifflin and Company, 1930. 310 pp.
- BARR, A. S., and RUDISILL, MABEL. "An Annotated Bibliography on the Methodology of Scientific Research as Applied to Education," *University of Wisconsin, Bulletin of the Bureau of Educational Research*, No. 13. Madison: University of Wisconsin, 1931. 129 pp.
- BIXLER, H. H. *Check Lists for Educational Research*. New York: Bureau of Publications, Teachers College, Columbia University, 1928. 118 pp.
- CRAWFORD, C. C. *The Technique of Research in Education*. Los Angeles: The University of Southern California, 1928. 320 pp.
- GOOD, C. V. *How to Do Research in Education*. Baltimore: Warwick and York, 1928. 298 pp.
- KELLEY, T. L. *Scientific Method; Its Function in Research and in Education*. New York: The Macmillan Company, 1932. 233 pp.
- MONROE, W. S., ODELL, C. W., HERRIOTT, M. E., ENGELHART, M. D., and HULL, M. R. "Ten Years of Educational Research, 1918-1927," *University of Illinois Bulletin*, Vol. 25, No. 51, Bureau of Educational Research Bulletin No. 42. Urbana: University of Illinois, 1928. 367 pp.
- RUSK, R. R. *Research in Education; an Introduction*. London: University of London Press, 1932. 108 pp.
- WAPLES, DOUGLAS, and TYLER, R. W. *Research Methods and Teachers' Problems*. New York: The Macmillan Company, 1930. 653 pp.
- WHITNEY, F. L. *Methods in Educational Research*. New York: D. Appleton and Company, 1931. 335 pp.

CHAPTER II

EDUCATIONAL PROBLEMS AND THEIR DEFINITION

The educational problems being studied. The educational problems to which research workers are giving their attention are indicated by the titles of reports appearing in such periodicals as *Journal of Educational Research*, *Journal of Educational Psychology*, *Educational Administration and Supervision*, and *School Review*. They are possibly more typically represented by lists of graduate theses in education. The following titles were taken at random from the 1929-1930 Bibliography of Research Studies published by the United States Office of Education.

1. A survey of the public schools of Imperial County, California
2. The educational ideas of Louisa May Alcott
3. The comparative influence of motion pictures in teaching American history
4. Analysis of integration; a study of the relationship between eye, hand, and foot response mechanisms
5. A study of the relation between developmental age and some physical measurements
6. A critical study of standardized mechanical aptitude tests
7. Bureaus of research in public school systems with reference to cities of 10,000 population or less
8. Kinesthetic factors in the learning of reading and spelling
9. A study of certain sound letter confusions in spelling in grades two to six
10. Relationship of reading ability and success in high school English in the junior class of the Milne high school
11. How literary artists of the nineteenth century were influenced by current psychology and philosophy in delineating children
12. What skills in mathematics are necessary in order that a student may do the mathematics required by some colleges in the first year of a course leading to a B.A. degree
13. Some number abilities of beginners in rural and town schools

14. A course in general science
15. An evaluation of certain standard tests in high school physics
16. Methods in the teaching of high school history
17. A study of the dominant characteristics of adolescent children having superior untrained musical talent
18. A study of the clothing weights and physical activity together with the possible correlation of these in the Merrill-Palmer nursery school
19. State high school standardization
20. Relationship of scores obtained by junior high school pupils in the Rogers physical fitness tests to their mental ability and achievement

Some of these titles are not very definitive but the list is indicative of the wide range of problems being studied in the field commonly designated as educational research. This field has no definite boundaries. It obviously overlaps that of psychology and sociology. There are occasional excursions into other fields. A topical index of approximately 3650 reports of educational research, and related materials, for the period 1918-1927, not including articles in periodicals, included 605 items exclusive of duplications. It may be unfortunate that educational research has not been confined to a more limited field, but a more restricted definition of the field would not be in conformity with prevailing practice.

A classification of the problems of education. The picture of the field of educational research afforded by the preceding paragraphs is general. Several writers have attempted a more meaningful description by listing types of educational research, but there is little agreement in the classifications proposed. The diversity of view is illustrated in "A Symposium on the Classification of Educational Research" published in the *Journal of Educational Research*, May and June, 1931. Whitney¹ has tabulated the rubrics in twelve classifications. No rubric of the twenty-two that appear in the table was recognized by all twelve of the authors and only two rubrics (experimental and survey) were listed by as many as eleven of the authors.

¹ Whitney, F. L. *Methods in Educational Research*. New York: D. Appleton and Company, 1931, pp. 72-73.

One explanation of this lack of agreement in these classifications is to be found in the different points of view from which analyses have been made.¹ Most writers have attempted differentiation on the basis of techniques employed. This approach leads to difficulty. It has been asserted that "all research involves the use of statistical methods."² A classification of problems according to the field of investigation results in a bewildering array of rubrics unless the fields³ are very broad and such an analysis is not very useful. For the purpose of this text, an analysis on the basis of the type of question or problem is proposed.⁴ Questions asking what has been (historical) and questions asking concerning the status of present conditions and practices (survey) are obvious as general types. A number of questions relate to the measurement of human abilities and traits. Studies of relationship are grouped under three heads. The determination of what should be or the answering of questions of value represents another type of problem, and the evaluation and summarization of research is listed as a final type.⁵

1. Historical
2. Measurement (Construction and validation of measuring instruments)
3. Survey (Determination of status of conditions and practices)
4. Experimental (Determination of relative effectiveness of comparable procedures)
5. Concomitant variation (Prediction)

¹ Freeman, F. N. "A Symposium on the Classification of Educational Research," *Journal of Educational Research*, 24: 16-19, June, 1931.

² Symonds, P. M. "A Course in the Technique of Educational Research," *Teachers College Record*, 29: 24-30, October, 1927.

³ Ayer has proposed fifteen fields of administrative research. Ayer, F. C. "Administrative Research in Public Administration," *The Nation's Schools*, 2: 14, September, 1928.

⁴ The organization of Chapters VI-XIII of this text conforms to this classification. The reader should consult these chapters for further explanation of the several rubrics.

⁵ The reader should note that there is no specific mention of problems relating to the development of statistical procedures. Most of such problems occur incidentally in dealing with the types designated by the second, fourth, fifth, and sixth rubrics.

6. Causal (Identification of causes and measurement of relative contributions)
7. Determination of values, or of what should be
8. Evaluation and summarization of research

Relative importance of the several rubrics of problems. If a group of superintendents, principals, and teachers were asked to list the problems that they consider important, a large proportion of the questions would belong under the head of survey research. The practical schoolman is interested in the status of his school system and of the several units within it. He desires to know concerning practices and conditions in other schools. Beyond such inquiries his interests are scattered. Sometimes he asks concerning the relative merits of alternative procedures and practices. Occasionally he is interested in causes of certain conditions and the effects of certain procedures. If, however, we consider the question of the relative importance from the point of view of a science of education which is conceived of as a systematized collection of general facts and principles, the answer will be different. Survey investigations, which are frequently characterized as fact-finding studies, will be assigned a place of minor importance.

The problems fundamental to a science of education.¹ Since a large portion of educational research involves the measurement of human abilities and traits, the problems of measurement are fundamental and the most fundamental are those having to do with the identification and specification of the abilities and traits whose measurement is desired. Some progress has been made toward determining what general intelligence tests measure and there is some information concerning the nature of achievement in the various subject-matter fields. For example, it has been shown that what many achievement tests measure includes as a major factor the same ability that is measured

¹ The problems noted here imply certain assumptions. For example, the problem of measurement assumes a sufficient degree of stability in human abilities and traits to make measures of them meaningful. The identification and formulation of such assumptions might be included as one of the problems fundamental to a science of education.

by typical tests of general intelligence. We commonly speak of mathematical ability but a recent investigation¹ indicates that there is no general mathematical ability. Several investigators have interpreted their data as indicating that ability in the field of arithmetical calculation is highly specific, but the total evidence is not consistent and the identification of the abilities that function in arithmetical calculation is still a problem. Another problem under this head involves the question of permanency of achievement. When a measure of achievement is secured it is conceived of in terms of future performance, but the date of this performance is rarely specified. We need to know how rapidly an ability will deteriorate during the period following the date of testing and how to predict the performance it will make possible at any specified date.

A second fundamental problem is the identification of causes. Questions under this head ask, "What are the causes of?" "Does.....contribute to.....?" "What are the factors that contribute to.....?" An important problem under this rubric is to determine the factors (instructional techniques, size of class, etc.) that affect learning.

A third problem may be expressed as follows: Given a relationship involving as causes such factors as instructional techniques, the personality of the teacher, size of class, and general organization of the school, to determine the effect upon the dependent variable of specified changes in a designated cause. The dependent variable is usually a measure of a specified segment of the achievement of a group of subjects. The mean of the measures of the achievement is most commonly used, but the dependent variable may be some other function of the distribution. The problem of the effect of training (practice) upon individual differences is an illustration of the use of a measure of variability. An important special case under this general problem is formed when the varying independent vari-

¹ Cairns, George J. "An Analytical Study of Mathematical Abilities," *Catholic University of America, Educational Research Monographs*, Vol. 6, No. 3. Washington, D. C.: Catholic Education Press, April, 1931. 104 pp.

able is time as in the constancy of the IQ and other genetic problems.

The measurement of the contributions of given causes furnishes a fourth fundamental problem. Closely related is the problem of determining the status of the various factors which constitutes the optimum conditions for engendering specified achievements. At present we have hypotheses concerning these optimum conditions. The advocates of the activity curriculum have proposed a general description of the conditions which they insist will result in maximum learning. Other groups of educational theorists proclaim confidence in other conditions. But our actual knowledge concerning optimum conditions for learning is exceedingly fragmentary and until we have a dependable determination, our study of many of the problems relating to the training and selection of teachers, and the organization and administration of our schools will not have a secure foundation.

The identification of the factors having a predictive value and the determination of the best basis for making desired predictions within specified populations furnishes a fifth fundamental problem.

The sixth fundamental problem is the determination of the objectives of education. The questions involved ask what should pupils learn or what pupil achievements should be desired. Some writers, especially those with training in the field of philosophy, contend that these questions are outside of the realm of educational research. It is true that the answers cannot be developed entirely from objective data, and this fact has been recognized by certain writers who list philosophical inquiry as a type of educational research. Whether philosophical research is or is not recognized as a type is not important, but it is important to realize that the determination of the goals of education is a fundamental problem. Until these goals are specified in terms of measurable pupil controls of conduct, any determination of optimum learning conditions, which include methods and materials of instruction and the organization and administration of our schools, will rest on assumptions or arbitrary postulates.

In this exposition of the problems fundamental to a science of education there has been no mention of the population for which they should be studied. With the exception of those relating to measurement, most of the questions asked should be studied for several populations. Hence, the reader should think of the fundamental problems noted as very general. In defining one of them a particular population should be specified.

Discovering problems for graduate theses in education. A large proportion of first-year graduate students and all candidates for the doctorate face the task of discovering and selecting a suitable thesis problem. The preceding reference to the fundamental problems of education may suggest to the reader that the graduate student should direct his attention to them and endeavor to formulate a problem whose solution will give promise of contributing to the development of a science of education. Unfortunately first-year students and most candidates for the doctorate do not have the resources and the time required for dealing with most of the fundamental problems, and hence in general such investigations must be left to experienced persons who have adequate resources and who may, if necessary, continue the research over a period of years. Hence, graduate students, especially candidates for the master's degree, are restricted to survey investigations or other types of inquiries that are commensurate with their training, resources, and time. With this general restriction in mind, we may consider how suitable problems may be discovered.

Difficulties encountered in the course of practical activities are a fruitful source of problems. Hence, it may appear that the graduate student who has taught or served in an administrative position should approach his thesis work with a number of problems in mind. It is, however, unusual for a student to begin his graduate study with a definite problem in mind. Several factors contribute to this condition. Many teachers and administrators are not sensitive to difficulties and consequently do not become aware of many potential sources of educational problems. Even when a difficulty has been recog-

nized, it is frequently not easy to identify the problem and to state it in satisfactory terms. Furthermore, practical problems are frequently difficult to deal with in a scholarly way. Hence, the practical experience of a graduate student is seldom a fruitful source of appropriate problems.

It is helpful to examine a number of graduate theses and other reports of research. Frequently an author suggests problems for study. Other questions may grow out of a critical reading of a report of educational research. The possibility of employing different techniques for the studying of the same problem may occur to the reader. Writings about research and research techniques frequently suggest problems for study. Some writers have discussed the need for research in particular fields.¹ Critically minded instructors frequently suggest problems in connection with their courses and the alert graduate student will usually be able to accumulate a list of problems from the courses he is taking. Term reports, especially summaries of the research relating to a given topic, usually suggest several problems for study.

The alert graduate student should encounter no difficulty in discovering a number of problems. Whenever a problem occurs to him it should be recorded. It is desirable to formulate it in concise terms, preferably in question form. A topical statement is seldom very definitive.

Selecting a problem for a graduate thesis.² The reader should note that the reference in this paragraph heading is to a "problem" rather than to a "topic." The student should think in terms of questions or groups of questions rather than topics. Until the problems being considered are conceived of in terms of questions to be answered, the student cannot wisely evaluate the possibilities.

It is frequently stated in the requirements of a graduate

¹ See bibliography at end of this chapter.

² Although the following reference is directed to another audience, a student selecting a problem for a graduate thesis may study it with profit.

Symonds, P. M. "Common Faults in Graduate Research in Education," *Journal of Educational Research*, 27: 481-92, March, 1934.

school, especially those which have been set up for the degree of doctor of philosophy in education, that the thesis shall represent an "original contribution." The term "original contribution" does not appear to possess any very precise meaning. Is a problem that has been studied thereby disqualified as a basis for an "original contribution"? The findings of many researches in the field of education, especially the earlier ones, are not dependable. Many of the problems studied are in need of reinvestigation with the utilization of more adequate techniques. It is the opinion of the present writers that very frequently problems which have been studied are appropriate for further research. An "old" problem may be considered suitable for a thesis, provided the student is able to apply improved techniques or to apply more skillfully procedures previously employed. In the case of a fundamental problem, the repetition of an investigation for the purpose of verification may be justified.

Some persons do not recognize the synthesis, interpretation, and application of the findings of other investigators as "original contributions." Sometimes this position is qualified by saying that such undertakings are acceptable as masters' theses.¹ The science of education is in need of evaluation, interpretation, and synthesis of the findings of educational research.² The evidence scattered throughout the research literature in support of hypotheses, rules, and principles needs to be critically examined and the dependable findings synthesized so that the justification for accepting these hypotheses, rules, and principles, may be established. There is also urgent need for the interpretation of this scattered evidence in terms of applicability to educational practice. It appears, therefore, that a critical, scholarly summary may be recognized as an "original contribution,"

¹ Several institutions accept for the doctor's degree theses which represent "organizations and applications of existing knowledge." This modification is more characteristic, however, of the institutions which grant the degree of doctor of education. See Monroe, W. S. "A Survey of the Requirements for the Doctor of Philosophy in Education," *School and Society*, 31: 655-61, May 17, 1930.

² A more extended treatment of this point is given in Chapter XIII.

provided the field of the summary has been sufficiently studied. The question of accepting a critical summary in partial fulfillment of the requirements for a graduate degree involves other considerations. To the present writers the acceptance of a critical summary does not seem inappropriate in the case of the master's degree.¹

The contributory value of a mere fact-gathering investigation is usually slight. A similar statement may be made with reference to studies whose significance is primarily local. In general studies of relationships have a higher contributory value than survey investigations. Furthermore, the possibility of making a contribution is greater when the study is based upon or is related to previous researches. Educational research has been characterized as fragmentary which means that the efforts of investigators have not been coördinated and that when a synthesis of findings is attempted they often are so unrelated that dependable generalization is not possible. In many cases the situation is made still more unsatisfactory by reason of the fact that a number of the investigations are superficial or have dealt with trivialities.

In addition to the possible contributory value of the research, the student should consider the educative opportunity it will afford him. He may properly regard the thesis requirement as a major learning exercise. If the work involved is largely routine and mechanical, the student will learn very little even if he succeeds in satisfying the institutional requirement.

The feasibility of the procedures for attacking the problem should also be considered. Many important problems are unsuitable for graduate theses because the indicated procedures are not feasible or needed instruments and techniques have not been developed. A phase of the feasibility of the indicated procedure is the time required to complete a proposed inquiry. It is characteristic of inexperienced persons to underestimate

¹ For discussion of certain aspects of the thesis requirement for the doctorate, see Buswell, G. T. "Research and the Degree of Doctor of Philosophy in Education," *Journal of Educational Research*, 23: 146-52, February, 1931.

the time element and a graduate student should seek the advice of an experienced person on this point. It is much better to study intensively a restricted problem than to spread out one's efforts over a larger problem with the result that the research deserves the characterization of superficial.

Finally, the student should consider his training and his resources. A problem may be inappropriate for a particular student because his training has not provided him with the necessary background. For example, a student should not attempt a curriculum problem if he has not taken one or more basic courses in this field. Neither should a student attempt a problem which will require intricate statistical techniques if he has not had adequate statistical training. A student should not attempt an experimental problem until he has attained a basic understanding of the experimental procedure. The student should also consider his resources for collecting the needed data and for handling them. Some problems require that the collection of data extend over two, three, or even more years. Collecting data for other problems requires considerable expense for test materials or for other purposes.

Defining a problem. In order to be effective as a guide in the subsequent phases of the research, a problem must be defined so that there will be a concise and complete statement of the specific question or questions to be answered. For example, consider the problem, "What is the relation between achievement in algebra and in Latin?" As stated this problem is very general. It should be restricted. One desirable restriction is with reference to the period during which the two subjects are studied. Another relates to whether the two subjects are studied simultaneously or in sequence. When these restrictions are incorporated, we have a statement of the following type: "What is the relation between achievement in first-year algebra and achievement in first-year Latin when the subjects are studied simultaneously in the ninth grade?" Further definition is needed with reference to the student group to be considered. As stated, the question applies to all students in all schools.

One basis of restriction in this respect is the requirement or non-requirement of the subjects being considered. Finally a definition of achievement is needed. When all of these items are incorporated the statement would appear as follows:

What is the relation between achievement (as defined) in first-year algebra and achievement (as defined) in first-year Latin when these subjects are studied simultaneously in the ninth grade in high schools which require algebra but make the study of a foreign language elective and offer one modern language in addition to Latin. Students repeating one or both of these subjects are to be excluded.

The problem may be further limited to public high schools and even to those of a certain size and within a certain area.

The definition of a problem should result in a precise and complete statement of the questions to be answered. Attention must be given to the terms used. In the illustration just given attention was called to the necessity of defining "achievement." "Grades received" might be used instead of achievement, but to do so would introduce a significant change in the problem. Similarly, there is a difference between "costs" and "expenditures"; "test scores," and "measures of achievement"; "intelligence as measured," and "intelligence." Frequently, the problem analyzes into a group of questions. Each of these should be stated or at least definitely indicated.¹

The function of the definition of a problem. Although a problem may be discovered when examining an accumulation of facts, the logical sequence is the problem first and then the collection of data to fit the problem. In this sequence the definition of the problem serves as a guide in determining what data should be secured and in collecting them. This function of the definition of a problem is more important in some cases than in others. In historical research, case studies, and a few other types of problems, a precisely defined problem may not be necessary because the purpose is to discover the implications of the data that it is possible to collect. In general, however, the researcher should endeavor to define his problem before

¹ For a discussion of defining experimental problems see Chapter IX.

the collection of data is begun and it is wise to reduce this definition to writing. Without a clearly defined problem, an investigator may waste time in collecting data. He may collect unnecessary data. He may fail to collect some essential data. It may happen that the data collected are not those called for by the problem.

The definition of the problem enables one to anticipate the statistical treatment of his data. Occasionally a graduate student or other investigator collects data and then finds that he does not have command of the statistical techniques required for dealing with them. A few years ago a graduate student who had only incidental training in statistical methods came to the senior author to ask how he should handle the data he collected. The problem, which had not been adequately defined, required the use of partial correlation. The student found the mastery of this technique too much of an undertaking and selected another problem for his thesis. The labor he had expended in collecting his data was wasted. Even persons of experience in educational research sometimes do not realize the difficulties in store for them until the problem is adequately defined. In reporting his study of the influence of nurture upon intelligence T. L. Kelley¹ confesses in the preface that although he had discussed the problem for years and had believed it to be a relatively simple one, he found that he had never clearly defined either "nurture" or "nature." He also found that the necessary statistical techniques had not been devised.

SELECTED BIBLIOGRAPHY

- ALEXANDER, CARTER. "Research in Educational Publicity. Outstanding Achievements and Needed Studies," *Teachers College Record*, 29: 479-87, March, 1928.
- BARR, A. S. "Needed Research in Classroom Supervision," *Peabody Journal of Education*, 5: 209-15, January, 1928.
- BRIGGS, T. H. *Curriculum Problems*. New York: The Macmillan Company, 1926. 138 pp.

¹ Kelley, T. L. *The Influence of Nurture upon Native Differences*. New York: The Macmillan Company, 1926, p. v.

- BRIGGS, T. H., et al. "Topics Proposed for Research in Secondary Education," *Journal of Educational Research*, 14: 67-75, June, 1926.
- DAVIS, R. A. "Research and the Schools," *Journal of Educational Research*, 26: 561-68, April, 1933.
- DOUGLASS, H. R. "The Contribution of Research to Secondary-School Curriculum Construction," *School and Society*, 32: 411-16, September 27, 1930.
- GOOD, C. V. "Cautions to the Beginner in Educational Research," *Journal of Educational Research*, 26: 302-04, December, 1932.
- GOOD, C. V. "Research in Secondary School Methods," *Journal of Educational Research*, 22: 9-30, June, 1930, especially pp. 26-30.
- HENMON, V. A. C. "Needed Research in the Field of Learning," *Journal of Educational Research*, 11: 313-21, May, 1925.
- JUDD, C. H. "Research in Elementary Education," *Journal of Educational Psychology*, 17: 217-25, April, 1926.
- JUDD, C. H. "The Place of Research in a Program of Curriculum Development," *Journal of Educational Research*, 17: 313-23, May, 1928.
- KELLY, F. J. "Scientific Method in College Administration and College Teaching," *School and Society*, 20: 390-96, September 27, 1924.
- KLAPPER, PAUL. "The Experimental Study of Education with Special Reference to the Elementary School," *Journal of Educational Research*, 12: 123-35, September, 1925.
- KULP, D. H. "Problems of Rural Education Demanding Sociological Research," *Teachers College Record*, 31: 332-38, January, 1930.
- LEE, E. A. "Research Problems in Training Vocational Teachers," *School and Society*, 24: 31-37, July 10, 1926.
- MARTIN, C. W. "Problems of Higher Education as Found in Periodical Literature," *Peabody Journal of Education*, 9: 372-76, May, 1932.
- MEAD, A. R. "List of Possible Studies and Researches in Supervised Student-Teaching," *Educational Administration and Supervision*, 11: 355-58, May, 1925.
- MONROE, W. S. "Service of Educational Research to School Administrators," *American School Board Journal*, 70: 37-39, 122, 125, April, 1925.
- MORT, P. R. "Needed Research in the Field of State Aid," *Teachers College Record*, 27: 707-12, April, 1926.
- NORTON, JOHN K. "Ph.D. and Ed.D. Adventures in School Administration," *Teachers College Record*, 36: 207-12, December, 1934.

- PAYNE, E. G. "Research Problems and Trends in Educational Sociology," *Journal of Educational Research*, 25: 239-52, April-May, 1932.
- PECHSTEIN, L. A. "The Problem of Negro Education in Northern and Border Cities," *Elementary School Journal*, 30: 192-99, November, 1929.
- SIMPSON, A. D. "Needed Researches in Public School Finance," *Journal of Proceedings and Addresses of the National Education Association*; Vol. 70. Washington: National Education Association, 1932, pp. 372-73.
- STENQUIST, J. L. "Getting Research into Practice in a Large School System," *American School Board Journal*, 79: 41-42, 131-32; November, December, 1929.
- STRANG, RUTH. "The Relation of the Dean of Women to Research," *Teachers College Record*, 31: 44-49, October, 1929.
- STRAYER, G. D. "The Scientific Approach to the Problems of Educational Administration," *School and Society*, 24: 685-95, December 4, 1926.
- SYMONDS, P. M. "Needed Research in Diagnosing Personality and Conduct," *Journal of Educational Research*, 24: 175-87, October, 1931.
- SYMONDS, P. M. "Needed Research in the Field of Measurement in Secondary Education," *Journal of Educational Research*, 16: 119-26, September, 1927.
- SYMONDS, P. M. "Needed Research in the Teaching of English," *English Journal*, 22: 447-56, June, 1933.
- THORNDIKE, E. L. "Curriculum Research," *School and Society*, 28: 569-76, November 10, 1928.
- THORNDIKE, E. L. "The Need of Fundamental Analysis of Methods of Teaching," *Elementary School Journal*, 30: 189-91, November, 1929.
- WATSON, G. B. "Needed Investigations in the Psychology of Character," *Religious Education*, 23: 66-72, January, 1928.
- WHEELER, H. E. "Suggestions for Research on the Typography of School Textbooks," *Elementary School Journal*, 29: 27-31, September, 1928.
- WOODY, CLIFFORD. "The Values of Educational Research to the Classroom Teacher," *Journal of Educational Research*, 26: 172-78, October, 1927.

CHAPTER III

COLLECTING THE DATA SPECIFIED BY A PROBLEM—BASIC TECHNIQUES

Educational research not a mechanized routine. Although relatively definite techniques will be described in this and several of the following chapters, the reader should not infer that educational research can be reduced to a mechanized routine. A problem, whose study is worthy of designation as educational research, represents a “new” situation. In meeting this situation the research worker must define the problem and then determine what technique, or techniques, are appropriate for collecting the necessary data. In making this decision he must consider not only the nature of the problem but also the conditions which he will likely encounter in collecting his data. For example, he may have to choose between two techniques on the basis of their relative feasibility with respect to the conditions current at the time of the investigation. After a technique has been determined upon, it is frequently necessary to make adaptations to the problem and to the conditions under which the research is conducted. Sometimes considerable ingenuity is required to obtain the needed data. The handling of the data collected may be a relatively routine activity but the interpretation of the derived statistics is likely to require critical reflective thinking of a high order. Hence, the reader of this text should, at all times, bear in mind that educational research cannot be conducted in a routine fashion.

Types of data. When reference is made to the data of educational research, one is likely to think of test scores, school marks, chronological ages, teachers’ salaries, counts of things such as pupils or language errors, and the like. The term, however, is used with a broader meaning. In dealing with some

educational problems it is necessary to secure statements of beliefs or opinions, statements descriptive of schools or events, conclusions reached by previous investigators, or historical information. Hence, data should be thought of as including all facts, concepts, and principles useful in deriving answers to the questions listed in the definition of a problem.

Data that represent measures or counts of things are expressed in numerical terms and are commonly referred to as quantitative. Statements of beliefs or opinions, descriptions of schools or events, and the like are referred to as non-quantitative data.

Objective data versus subjective data. The data of educational research are often regarded as subjective when they consist of expressions of judgment made by subjects participating in the investigation as sources of data. From this point of view, responses to many questionnaires would be labeled "subjective data." It is more desirable to regard as subjective, data obtained in such a manner that they may be influenced by the person collecting them. For example, observations of the behavior of small children made by a research worker would be subjective because, in collecting the data, he gives meaning to his sensations in perceiving overt acts. To the extent that another research worker making the same observations would be likely to give different meanings to the "sense data," we are justified in regarding the recorded observations as subjective.

The converse of this statement gives the definition of objectivity which will be used in this book. When data are such that there has been very little or no opportunity for them to be affected by the prejudices, opinions, and judgment of the person collecting them, they are to be regarded as objective.

It should be noted that the subjectivity of data is a matter of degree. Certain types of data are regarded as "highly subjective" because experience has shown that in general two persons working independently will obtain significantly different data. This is true in the case of marks assigned to examination papers of the essay type. Other types of data are known to be

only slightly influenced by the person who collects them and may be described as "slightly subjective." Sometimes such data are designated as "objective" but it is wise to reserve this term for data whose degree of subjectivity approaches zero. In other words, "subjective" has a somewhat broad and relative meaning while "objective" has a more restricted meaning.

The problem a guide in collecting data. The problem, when adequately defined, specifies the data to be collected, and usually the items collected should be limited to the requirements of the problem. Collecting unnecessary data consumes time and may distract the attention of the investigator from the issues of his problem. Occasionally it may appear desirable to redefine the problem so as to extend its scope, or the investigator may find that it is feasible to collect data for one or more related problems without adding greatly to the time required for this phase of the work. It is unwise, however, to collect data merely because they are thought to be interesting. Data which are not used represent wasted effort.

The basic techniques for collecting data. Although the details of the procedure followed in a given problem are determined by its nature and the availability of sources and instruments, there appear to be only a few basic techniques. In the following exposition, seven are recognized.¹

1. Copying data from records and published sources.
2. Analyzing texts, records of activities, etc.
3. Interviewing.
4. Constructing and using questionnaires.
5. Securing a current record of activities, etc.
6. Making estimates and ratings.
7. Selecting and administering tests.

1. *Copying data from records and published sources.* When data are to be copied from records or published sources, the principal considerations are accuracy and economy of time.

¹ Locating published sources of information and collecting data of the sort that are utilized in dealing with philosophical problems or in preparing a summary of research are dealt with in Chapters XII and XIII.

Accuracy in copying is attained when the entries made are identical with those in the source. The number of errors can be reduced to a minimum by employing certain techniques and devices, but checking is necessary to insure accuracy.

The details of the process of copying and the form in which the copied items should be arranged vary with the nature of the source and the use that is to be made of the data. A resourceful person will usually think of devices that will facilitate the work as well as reduce the errors of copying to a minimum. When data are being copied from several columns of a table, a strip of cardboard may be used to assist in following a line across the table. Alexander¹ has suggested a helpful device when data are to be copied from only certain columns of a table. It consists of a strip of paper, or cardboard, one edge of which is notched so that when it is placed horizontally on the table, the items to be copied are exposed and the irrelevant items are masked.

In copying numerical data from records and published sources it is important to label accurately each item or each group of similar items. What this means may be illustrated by reference to numerical data recorded in a publication of the United States Office of Education.² For example, Table 10 of this publication gives data pertaining to students in state normal schools.³ Any item copied from this table should be accompanied by the words, or phrases, which give meaning to the datum. For example, one figure given in the table may be copied "2805 women resident state normal school students enrolled in regular session in New Jersey, 1927-28." The fact that this figure includes both white and colored women students should probably be recorded also. The figure 7291 when copied should be given the label: "resident men students in all courses, excluding

¹ Alexander, Carter. *School Statistics and Publicity*. Boston: Silver, Burdett and Company, 1919, p. 85.

² Phillips, F. M. "Statistics of Teachers Colleges and Normal Schools, 1927-28," *United States Bureau of Education Bulletin*, 1929. No. 14. Washington: Government Printing Office, 1929. 71 pp.

³ *Ibid.*, p. 19.

duplicates, enrolled in state normal schools of continental United States, 1927-28; including both white and colored men students." If the precise meaning of the label is not apparent, it should be ascertained and recorded. Such labels as "per pupil in average daily attendance," "per cent," and even "enrollment" may vary in meaning.

Economy of time in copying is attained by devising an appropriate plan of work and by employing special devices such as the one described by Alexander. Contributions are made to economy in the total time of the research by copying the data in a form that will facilitate their tabulation and by making certain that all of the needed data are copied when the source is at hand.¹ Frequently much time can be saved by anticipating the tabulations to be made from the copied data. Items copied from several sources which pertain to the same topic can be copied on the same card, or page. Items can be grouped so that calculation of measures of central tendency, such as means, is greatly facilitated. Occasionally, it may be desirable to plan the tables which are to appear in the reported research and to organize the copying of numerical data so that blanks in the prepared tables may easily be filled in. Careful planning reduces the probability of omitting needed items in copying. Furthermore, careful planning will tend to restrict the tendency to copy interesting but irrelevant data.

Although it is not a part of the process of copying, the investigator should inquire concerning the accuracy of the sources. Sometimes one source may be checked against another. An illustration of this is found in the work of Rugg in the St. Louis Survey. In this investigation Rugg checked the statistics of the United States Bureau of Education against the statistics of the United States Bureau of Census and showed, for example, that "the per pupil cost for salaries and expenses of supervisors, principals, salaries of teachers, repairs, textbooks, salaries of janitors, obtained from the two sources . . . show a very

¹ Failure to record an adequate bibliographical reference is not an uncommon fault of inexperienced investigators.

satisfactory agreement. . . ."¹ The investigator who utilizes numerical data from records and published sources should regard his data in somewhat the same manner as the historical research worker. He should seek to establish the validity and accuracy of his data by attempting to secure "affirmations of independent witnesses."

Finally, when copying data from records and published sources, it is important to make a complete record of the exact source from which the data are being taken. Such a record serves three purposes: First, the source may be more easily located again if further data are needed. Second, the source may be more easily located for checking the accuracy of the data copied. Third, it is very desirable to present in the report of the investigation adequate and accurate citations of the sources of the numerical data. It is not scholarly to do otherwise.

2. *Analyzing textbooks, courses of study, records of activities, and the like.* In analyzing textbooks, the investigator may wish to determine what topics are treated, what words are used, what types of learning exercises are given, and so on, or his purpose may be that of determining the amount of space given to topics, the frequency and order of appearance of words, the frequency of different types of learning exercises, and the like. In analyzing courses of study he may wish to determine what courses are offered and the characteristics of these courses. Pupil writings may be analyzed with respect to language errors.

The identification of words, language errors, topics, and other characteristics requires the recognition of certain criteria. The problem, when adequately defined, specifies or at least implies the criteria to be observed. For example, if the problem calls for the determination of the arithmetical processes needed in solving problems given in commonly used chemistry texts, the meaning of the term "arithmetical processes" furnishes a

¹ Rugg, H. O. "Public School Costs in St. Louis," *Survey of the St. Louis Public Schools*, Part III—Finances. Yonkers-on-Hudson, New York: World Book Company, 1918, p. 21.

criterion. More precise definition may specify the identification of particular number combinations.

The analysis usually requires careful reading of the texts, courses of study, or other materials during which the items that satisfy one or more of the criteria are identified and copied on data sheets or cards. Occasionally the reading may be completed before the work of recording is begun. In such cases the investigator may check with suitable symbols the items that are to be copied later. This is an excellent procedure when a trained analyst is needed in the identification of the items, but the work of copying may be left to clerks.¹

The criteria to be observed may be simple. For example, if pupil compositions are being analyzed for misspelled words, the spellings given by a dictionary furnish the criterion.² When the problem is to determine the "vocabulary load" of textbooks, the Thorndike *Word Book*³ is often used as the criterion. This list contains Thorndike's determination of the ten thousand most commonly used words in the English language and when used as a criterion the analysis becomes the identification of the words not found in this list. The analysis may be made

¹ The following references may be consulted for further discussions of the techniques used in analyzing textbooks:

Fowlkes, J. G. *Evaluation of School Textbooks*. New York: Silver Burdett and Company, 1923. 33 pp.

Franzen, R. H., and Knight, F. B. *Textbook Selection*. Baltimore: Warwick and York, 1922. 94 pp.

Hall-Quest, A. L. *The Textbook: How to Use and Judge It*. New York: The Macmillan Company, 1918. 265 pp.

Maxwell, C. R. *The Selection of Textbooks*. Boston: Houghton Mifflin Company, 1921. 138 pp.

Spaulding, F. E. *Measuring Textbooks*. New York: Newson and Company, 1922. 40 pp.

² There will be need for a few supplementary rules to furnish a basis for distinguishing between misspellings and errors of diction or grammatical errors.

³ Thorndike, E. L. *The Teacher's Word Book*. New York: Bureau of Publications, Teachers College, Columbia University, 1921. 134 pp.

The use of this criterion is illustrated in the following references:

Lively, B. A., and Pressey, S. L. "A Method for Measuring the Vocabulary Burden of Textbooks," *Educational Administration and Supervision*, 9: 389-98, October, 1923.

Remmers, H. H., and Grant, A. "The Vocabulary Load of Certain Secondary School Mathematics Textbooks," *Journal of Educational Research*, 18: 203-10, October, 1928.

more detailed by identifying the words in the first three thousand, those in the next two thousand, and so on. Thorndike's list is not entirely satisfactory as a criterion because no attention is given to the particular meaning with which a word is used.¹ Precise identification of technical terms in a particular field would require a more complex criterion.

The more common analyses of records of activities² are those of pupil writings for language errors or misspelled words and of arithmetical work for calculation errors. Charters³ has reported an attempt to ascertain the arithmetical operations used by salespeople. He selected at random 7337 charge checks (records of purchase transactions in which the goods are charged to customer accounts) and analyzed them for the addition and multiplication combinations involved. Another illustration is the analysis made of the *Reader's Guide to Periodical Literature* for the three-year period of 1919-21.⁴ A card was made for each of the eleven thousand topics appearing in the Index. The number of articles relating to each topic was noted. The cards were then sorted into piles, "one pile for each general field of human action or interest that seemed to be indicated or called for by the cards themselves." The sorting was carried on in this way until an apparently satisfactory grouping was attained. This work of classification resulted in a list of 46 topics with a range of from 9920 articles on the topic of government down to 89 for mathematics. The research of Finley and Caldwell⁵ illustrates another type of analysis. They sought to determine the extent to which biological material appears in the public press.

¹ For a discussion of the importance of meanings in vocabulary analysis see Dolch, E. W. *Reading and Word Meanings*. Boston: Ginn and Company, 1927. 129 pp.

² Records of activities include not only the types referred here, but also those secured by motion picture cameras, sound recording instruments, and other apparatus, and by observational procedures. See page 47.

³ Charters, W. W. *Curriculum Construction*. New York: The Macmillan Company, 1923, pp. 231-36.

⁴ Bobbitt, Franklin, et al. "Curriculum Investigations," *Supplementary Educational Monographs*, No. 31. Chicago: University of Chicago, 1926, pp. 7-22.

⁵ Finley, C. W., and Caldwell, O. W. *Biology in the Public Press*. New York: The Lincoln School of Teachers College, Columbia University, 1923. 151 pp.

Sometimes the formulation of criteria creates an important subordinate problem because the investigator is exploring a new field. For example, in a study of arithmetic texts by Monroe and Clark ¹ it was necessary to determine the elemental types of problems before the analysis could be begun and a list of 333 problem types was evolved only after several weeks of labor. In this case the task was unusually difficult because little attention had been given to the determination of problem types.

Many of the simpler types of analysis tend to be objective and a careful worker will secure highly accurate data, but not infrequently the process of analysis is rather highly subjective. In such cases steps should be taken to reduce the effect of subjectivity to a minimum. The first few hours devoted to the analysis should be considered a period of training and the work should be done over, preferably toward the close of the total period of analysis. A memorandum should be kept of all decisions made, so that consistency may be attained. All of the materials should be analyzed by the same person rather than be divided among two or more workers. If the materials can be analyzed independently by two or more persons, their results may be averaged.

3. *Interviewing.*² The interview and the questionnaire are used for securing unrecorded data in the possession of other persons. Both of these techniques involve subjective elements,

¹ Monroe, W. S., and Clark, J. A. "The Teacher's Responsibility for Devising Learning Exercises in Arithmetic," *University of Illinois Bulletin*, Vol. 23, No. 41, *Bureau of Educational Research Bulletin*, No. 31. Urbana: University of Illinois, 1926. 92 pp.

² The interested reader will find a more extended consideration in the following references:

Bixler, H. H. *Check Lists for Educational Research*. New York: Bureau of Publications, Teachers College, Columbia University, 1928, pp. 38-40.

Bogardus, E. S. "The Social Research Interview," *Journal of Applied Sociology*, 10: 69-82, September-October, 1925.

Bogardus, E. S. *The New Social Research*. Los Angeles: Jesse Ray Miller, 1926, pp. 69-130.

Sturtevant, S. M., and Hayes, H. "The Use of the Interview in Advisory Work," *Teachers College Record*, 28: 551-62, February, 1927.

Waples, Douglas, and Tyler, R. W. *Research Methods and Teachers' Problems*. New York: The Macmillan Company, 1930, pp. 519-32.

but when appropriate precautions are taken, the data obtained will usually be at least reasonably satisfactory.

In order to be successful the interviewer must be prepared for his task. An important phase of this preparation is determining the questions to be asked. An essential characteristic of the questions formulated is their relevance to the problem. The words and phrases used in asking the questions should be such as are likely to be understood by the individuals interviewed. It is desirable to avoid "leading" questions, since they tend to bias the responses. A question beginning "Do you—" will secure information of one type while a question commencing "How do you—" will secure a different type of information. Questions beginning "Do you—" and "How do you—" are instrumental in securing factual information, while questions beginning "Do you favor—" "How do you feel about—," "Do you recommend—" function in the collection of opinions, judgments, and the like. A question beginning "Why do you—" may secure either facts, or opinions, or both.

General questions are not likely to be very effective. For example, consider a typical group of elementary school teachers being asked, "What difficulties do you encounter in teaching silent reading?" Many of those interviewed will be able to recall only a few difficulties and probably none of them will mention all of their difficulties. More satisfactory results will be obtained if the interviewer asks specific questions: "Are you able to interest all pupils in silent reading?" "Are you able to determine the reading difficulties of pupils?" "Are you able to stimulate reading for enjoyment?"

The persons to be interviewed should be representative of the population or conditions for which conclusions are desired. Some steps should be taken to enlist the coöperation of these individuals. It is helpful in this connection to secure the sponsorship of some institution or association which those interviewed are likely to respect. If one wishes to interview the teachers in a given school system, the approach should be through their superintendent. The investigator should arrange

appointments that will be convenient to the persons interviewed, and which are for a sufficient duration of time. It is desirable to stimulate an appropriate frame of mind on the part of the person to be interviewed. This may often be done through an appeal to pride, some reference being made to the competence of the person to give the information desired. It is frequently effective to point out that the person interviewed is being given an opportunity to aid in the study of an important problem and that the information requested may not be obtained in any other way. The interest of the person interviewed can be stimulated by the interest exhibited by the research worker in his problem.

An interview should not degenerate into a quiz in which the person being questioned is caused embarrassment because he is unable to give the information desired. Personal or confidential matters should not be inquired into unless the interviewee indicates a willingness to be questioned about them. Whenever publication of the solicited information might cause embarrassment to the interviewee, assurance should be given that he will not be quoted and that information given in confidence will not be revealed.

It may not be wise to have a list of questions in evidence during the interview. It will sometimes be desirable for the investigator to memorize his list of questions. Usually, however, in collecting facts or opinions from mature individuals the presence of a question list should not result unfavorably; rather, it should aid in maintaining the control of the interview needed if all questions are to be answered. When it is evident that a question is not fully understood, the investigator should explain the terms used, or restate the question. If the investigator does not fully understand the response he should ask further questions. The investigator should be careful at all times to keep to the subject.

It is desirable to record in detail the responses of the subjects interviewed, but sometimes it may not be wise to do this in the presence of the subject. The time required for making

the record may interfere with the interview, or it may be that the subject will be intimidated or that a hostile attitude will be engendered. When it appears that either of these conditions will prevail, the interview should be written up as soon as possible after it has been made. The writers are of the opinion that most adults are not likely to be embarrassed by the open recording of their responses to questions. A superintendent, principal, or teacher is likely to respect the interviewer for his evident system. It is probable that the presence of a list of questions and the open recording of answers is most likely to have unfavorable effects when children are being interviewed.

The use of the interview is illustrated by the studies of Rufi¹ and Sharman.² In addition to securing data by the administration of tests, examination of records, and observation, Rufi collected many facts by interviewing principals and teachers in five small high schools. It is possible that some of these data could have been collected by questionnaire rather than by interview, but it is likely that this procedure would not have been as satisfactory. Some of the facts obtained by interview appear to be of the type which individuals would be hesitant to record on a questionnaire, but would be less hesitant to reveal in the course of a conversation. It is probable that in many cases the principals and teachers elaborated their replies so that they were clearly understood, which is not always true in the case of responses to a questionnaire. It is also probable that much information was volunteered relative to conditions, which the investigator would not have had the forethought to ask for if he had restricted his technique to the use of a questionnaire.

Data were obtained in the investigation of Sharman by interviewing principals, athletic coaches, physical education

¹ Rufi, John. "The Small High School," *Teachers College, Columbia University, Contributions to Education*, No. 236. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 145 pp.

² Sharman, J. R. "Physical Educational Facilities for the Public Accredited High Schools of Alabama," *Teachers College, Columbia University, Contributions to Education*, No. 408. New York: Bureau of Publications, Teachers College, Columbia University, 1930. 78 pp.

teachers, and other members of the faculties of a random sample of 38 per cent of the 275 public accredited high schools of Alabama. The investigator spent from one and one-half to two and one-half hours in each of the schools. Data were obtained relative to the distribution of full-time and part-time physical education teachers, physical examiners, requirements relative to health education, credit for physical education, types of activities in physical education programs, and facilities for physical education.

4. *Constructing and using questionnaires.* The questionnaire has long been subjected to criticism as a means of collecting data. The following comment was made in 1839.

It is impossible to expect accuracy in returns obtained by circulars, various constructions being put upon the same question by different individuals who consequently classify their replies upon various principles.¹

Thorndike made the following statement in 1911.

One vice of statistical studies in education today is the indiscriminate use of lists of questions as a means of collecting data by correspondence.²

Since 1911, numerous writers have criticized the questionnaire as a means of collecting data.³ Examination of the various criticisms reveals three general claims.

¹ "Report of a Committee of the Manchester Statistical Society on the State of Education in the County of Rutland in the Year 1838," *Journal of the Statistical Society of London*, 2: 303, October, 1839.

² Thorndike, E. L. "Quantitative Investigations in Education," *School Review Monograph*, Vol. 1, 1911, p. 43.

³ The following are some of the more extreme criticisms.

Burk, Frederic. "On a Certain Questionnaire," *School and Society*, 15: 170-73, February 11, 1922.

Butterfield, E. W. "Professional Ethics and the Questionnaire," *School and Society*, 11: 55-56, January 10, 1920.

Butterfield, E. W. "Educational Surveys," *Educational Review*, 68: 1-5, June, 1924.

Butterfield, E. W. "The Plenary Inspiration of the Dotted Line," *Educational Review*, 69: 1-4, January, 1925.

Flaccus, Quintus H. (Pseudonym). "Research in the Payment of School Executives," *School and Society*, 32: 806-07, December 13, 1930.

Hankinson, Frank. "The Blight of the Questionnaire," *Educational Review*, 73: 102-08, February, 1927.

(Continued next page)

1. The questionnaire is not an effective means of collecting data. The reasons given include (a) construction of questionnaire unsatisfactory, (b) responses carelessly made or mere opinions, (c) per cent of returns too low or the returns are not from a representative sample of the population.
2. The problem studied is unimportant, or the investigation is so limited in scope that the findings have little significance.
3. The questionnaire is an unjustifiable nuisance to superintendents, high school principals, and other recipients. Too many questionnaires are mailed out; many of them are unreasonably long; in many cases some of the information asked for is too personal or could be obtained from published sources; in some cases the recipient is asked to make time-consuming calculations.

When these claims are examined critically it is apparent that competent and courteous use of the questionnaire in the study of worth while problems is not condemned. The questionnaire is an important labor-saving device and it is obvious that its use as an instrument for collecting data cannot be completely dispensed with. The United States Office of Education, state departments of education, and other administrative units must use it as a means of collecting information that is generally recognized as valuable.¹ The cost of collecting this information by actual visitation and interview would be prohibitive. It would also be prohibitive in many survey investigations which cover a large area.

Ruckmick, C. A. "The Uses and Abuses of the Questionnaire Procedure," *Journal of Applied Psychology*, 14: 32-41, February, 1930.

Whitney, F. P. "Questionnaire Craze," *Educational Review*, 68: 139-40, October, 1924.

¹ For discussions from this point of view see

Bawden, W. T. "Searching after the Truth," *Industrial Education Magazine*, 27: 172-73, December, 1925.

Douglass, H. R. "The Questionnaire—To Be or Not to Be," *School and Society*, 15: 397-99, April 8, 1922.

Fallon, J. F. "The Questionnaire in Educational Research," *Catholic Educational Review*, 23: 539-45, November, 1926.

Koos, L. V. *The Questionnaire*. New York: The Macmillan Company, 1928. 178 pp.

Norton, J. K. "The Questionnaire," *Research Bulletin of the National Education Association*, Vol. 8, No. 1. Washington: National Education Association, 1930. 51 pp. A scale for rating questionnaires is suggested.

Perry, H. E. "The Questionnaire Method," *Journal of Applied Sociology*, 10: 155-58, November-December, 1925.

Although it is apparent that unqualified condemnation of the questionnaire is not justified, the present writers are fully aware that, even in the hands of a competent investigator, this instrument may yield faulty data. Wylie¹ has reported a personal first-hand investigation of the data secured by a questionnaire administered to eleven- and twelve-year-old pupils. Although the questions called for simple factual information, many answers involved a significant error. Self-interest of the respondents is sometimes a cause of a systematic error.² Mathews has shown that the order of the printed response words to be underlined on an interest questionnaire may introduce a systematic error.³ Frequently it is not possible to give precise and accurate answers to questions calling for factual statements descriptive of practices or conditions without adding much explanation. It is well known that failure of many of the recipients to respond to a questionnaire is a frequent cause of systematic errors in questionnaire data. Possibly the most adverse criticism that may be made of this technique, as it is most usually employed, is that the degree of reliability, validity, and representativeness of the data is uncertain.

The use of the questionnaire should be limited to worth while inquiries for which the needed data cannot be conveniently obtained by other means.⁴ Even in such cases the study should not be attempted unless the resources at the command of the investigator are such that success may reasonably be anticipated. These statements may seem platitudinous, but a sur-

¹ Wylie, A. T. "To What Extent May We Rely upon the Answers to a School Questionnaire?" *The Journal of Educational Method*, 6: 252-57, February, 1927.

² For a further discussion of this, see Stoke, S. M., and Lehman, H. C. "The Influence of Self-Interest upon Questionnaire Replies," *School and Society*, 32: 435-38, September, 27, 1930.

³ Mathews, C. O. "The Effect of the Order of Printed Response Words on an Interest Questionnaire," *Journal of Educational Psychology*, 20: 128-34, February, 1929.

⁴ Sometimes the desired data may be secured from published reports, and in other cases a modification of the problem will change the demands for data so that they may be secured from such sources. Institutional records and the files of state departments of education and of accrediting agencies are valuable sources of data that should not be overlooked.

prising number of questionnaire investigations deserve to be classified as of minor, if not trivial importance.

A questionnaire will be most successful when it is limited to requests for simple factual information in the possession of the recipients or easily accessible to them. A questionnaire that requires the recipient to spend much time in collecting or computing the requested information is not likely to be very successful and should be used only when the problem is one of considerable importance. A questionnaire asking for expressions of opinion should be employed only when the problem is important and the correspondents, in addition to being persons whose opinions will be significant, may be expected to be sufficiently interested so that they will give the questions thoughtful consideration.

The questions should be stated so clearly that they cannot be misinterpreted. Technical words and other unusual terms should be explained. The questions should be formulated so that they will not suggest or bias the answers. They should call for responses involving as little writing as possible. Questions calling for numerical data or for responses of "yes" or "no," underlining, or checking are most desirable. In some cases, questions should be inserted which will serve as checks to other questions. When this precaution is observed, the investigator is able to make an estimate of the reliability of his questionnaire data on the basis of "internal evidence."

The respondent should not be asked to do work which the investigator can do himself, such as computing totals, averages, or per cents. The questionnaire should be just long enough to secure the necessary data. Data irrelevant to the problem should never be requested. The investigator, however, should make certain that he has anticipated all of his needs. If the data required for the problem are not thus anticipated, the investigator may find it necessary to do without important information, or to impose a second questionnaire on his respondents.

In preparing a questionnaire considerable attention should

be given to its appearance. An attractive looking blank is more likely to secure a good response than an unattractive one. If possible, the questionnaire should be printed. When it is more than two or three typewritten pages in length and several hundred copies are required, printing is usually less expensive than mimeographing. The questionnaire should have a heading which includes an institutional or associational designation, and possibly a title. Spaces should be provided for the name, address, and position of the respondent. The questions should be spaced with care, ample room being allowed for responses. Respondents are less likely to overlook items when they are spaced effectively. The size of the questionnaire should be one that is convenient for handling and filing. Letter size (8½ by 11) is recommended. Legal size sheets are frequently inconvenient.

The first draft of the questionnaire should be submitted to competent and disinterested persons for criticism. If possible it should be given a preliminary trial by submitting it to persons typical of the proposed mailing list. Criticism and trial often reveal inadequacies not apparent to the author of the questionnaire. For example, it may be found that certain questions are easily misinterpreted. Analysis of these questions may disclose that the terms used are more technical than necessary or are in need of definition. Attempts to tabulate the data secured in a trial may suggest changes which will make for greater ease in tabulating.

The questionnaire should be accompanied by a tactful letter of explanation. The recipient should be informed of the nature of the problem and of its importance. It should be indicated that the responses will be held in confidence if their nature makes this action desirable. In this connection it should be mentioned that the questions should not be unduly personal in nature. Before sending out an elaborate questionnaire the willingness of the prospective recipients to respond should be determined by a preliminary inquiry. When this precaution is taken, greater interest is likely to be stimulated. Furthermore,

some of the expense of sending elaborate questionnaires to individuals disinclined to respond will be avoided. The recipient should be told that a summary of the data collected will be mailed to him. It is an excellent procedure in cases where the summary refers specifically to institutions or school systems, to accompany the summary with requests for criticism by its recipients. When these criticisms have been obtained, they may be used to improve the report of the study.

The questionnaire should be mailed at an opportune time. Seasons of vacation or periods of excessive activity should be avoided. For example, it is unwise to mail out a questionnaire just prior to the Christmas holidays or at the beginning or close of a semester. The questionnaire should be accompanied by a self-addressed stamped envelope for its return. A follow-up letter may be sent, under certain conditions, to those who fail to respond. Frequently, a tactful follow-up letter will greatly increase the per cent of responses. It is desirable to keep a record of the responses. In some cases it may be worth while to graph the number of questionnaires returned daily and to mail the follow-up letters when the curve falls unduly. Toops has reported two studies which reveal that the per cent of response may be increased to an acceptable figure by means of follow-up letters.¹ A similar conclusion has been reported by Lindsay.²

In connection with emphasizing what an investigator who employs a questionnaire should do, it may be pointed out that a recipient should recognize his responsibility. Many problems cannot be studied except by employing a questionnaire and when the problem is a worthy one and the investigator has exhibited competence and good judgment in preparing the questionnaire, the recipient should make an effort to respond. It is not courteous to throw it in the waste-basket. If there are

¹ Toops, H. A. "Returns from Follow-Up Letters to Questionnaires," *Journal of Applied Psychology*, 10: 92-101, March, 1926.

Toops, H. A. "Validating the Questionnaire Method," *Journal of Personnel Research*, 2: 153-69, August-September, 1923.

² Lindsay, E. E. "Questionnaires and Follow-Up Letters," *Pedagogical Seminary*, 28: 303-07, September, 1921.

valid reasons why the information should not be given, the questionnaire should be returned with a note of explanation. In answering the questions asked, an effort should be made to give correct information. If a question is not understood or if the recipient is unable to give an answer that he believes to be correct, the question should be omitted. When there is some doubt in regard to the meaning of the question asked, a notation indicating the interpretation given it may be made.

In concluding this discussion of the details of the questionnaire procedure, it may be reiterated that the problem should be worth while. The use of a questionnaire is not justified if it is feasible to collect the data by other means. The questionnaire should be carefully prepared and mailed to a wisely selected list of persons who are in a position to provide the desired information. The graduate student about to collect data by means of a questionnaire should secure the sponsorship of his institution. In many cases the sponsorship of an accrediting agency, or educational association, will secure an excellent response. Frequently, the letter accompanying the questionnaire should be signed by the sponsor rather than by the student. Many persons are unwilling to respond to a questionnaire from an unknown person, but are glad to respond to one authorized by an institution or organization that they respect. It should be mentioned in this connection that institutions and organizations should not violate this trust. The sponsor should accept the responsibility for making certain that the problem is worth while, a questionnaire necessary, and that the blank is well constructed. Furthermore, the sponsor should see that the respondents receive a summary of the results without undue lapse of time.

5. *Securing a current record of activity.*¹ When an activity does not normally result in an observable record, the various devices and techniques employed to secure a current record may be classified under three heads: (1) mechanical, (2) stenographic, and (3) observational. The moving picture

¹ The reader should bear in mind that after a record has been secured, analysis is required before the process of collecting data is complete.

camera, dictaphone, and other recording instruments are included under the first head. Usually the operation of the instrument requires little technical skill, but not infrequently considerable ingenuity is required to devise an instrument adapted to the conditions under which the record must be secured. The apparatus now used at the University of Chicago to secure a photographic record of the eye-movements of a reader is a good illustration.¹

When a competent stenographer or stenotypist is not available, a satisfactory record may be obtained by a person who takes a few notes and then soon afterwards expands them into a partial verbatim account. This method is generally employed by competent newspaper reporters, especially when interviewing persons whose time is limited. In an investigation reported by Helseth² a record of recitations was obtained by having two or more observers take such notes as they could. These notes were turned over to the teacher who expanded them into an approximately verbatim account of the recitation. Other cases have been reported by Horn,³ and Knudsen,⁴ and Stevens.⁵

The record secured by an experienced and competent stenographer may be expected to be highly accurate, at least so far as the meaning of the record is concerned, but if details of diction are considered, a stenographic record will usually involve errors.⁶ Obviously, a larger number of errors is to be expected when the record is an elaboration of notes taken by an observer. For some purposes, however, accuracy in the details of a record is not essential.

¹ See page 3 for references.

² Helseth, Inga Olla. "Children's Thinking," *Teachers College, Columbia University, Contributions to Education*, No. 209. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 163 pp.

³ Horn, Ernest. "Stenographic Reports of Speyer School Lessons," *Teachers College Record*, 16: 33-40, January, 1915.

⁴ Knudsen, C. W. *Evaluation and Improvement of Teaching*. Garden City, New York: Doubleday, Doran and Company, 1932, pp. 489-524.

⁵ Stevens, Romiett. "The Question as a Measure of Efficiency in Instruction," *Teachers College, Columbia University, Contributions to Education*, No. 48. New York: Bureau of Publications, Teachers College, Columbia University, 1912. 95 pp.

⁶ Greene, H. A., and Betts, E. A. "A New Technique for the Study of Oral-Language Activities," *The Elementary School Journal*, 33: 753-61, June, 1933.

In descriptive recording, the human element is even more important. An observer tends to interpret what he observes. For example, an observer may see that certain symbols appear on the blackboard. He perceives, or observes, that these symbols are addition examples. His perception or observation has an inferential element since he has given meaning to his sensations. Further inference may be made before the observation is recorded. He may infer that the class present in the classroom has been engaged in drill in addition, and further, he may formulate and record some inferences with respect to the quality of this drill. It is evident, therefore, that the skill of the observer and his experience are important factors in descriptive recording.

The observer must identify what he should record. Checklists, score cards, and other devices of this type are useful aids since they focus the attention of the observer on the items which are to be observed. Rating scales are an aid since they serve to control in some measure inferences of the evaluative type. Without the aid of some device to focus his attention, an observer, especially an untrained one, is not likely to report a satisfactory description. The record is likely to be a mixture of statements descriptive of action or conditions observed, inferences relative to mental activity, and evaluations of aspects of the scene observed.¹ Given a checklist or other suitable device, the record of a trained observer is likely to be rather highly reliable. An untrained observer working without a checklist or other aid is not likely to secure a reliable record unless the items to be recorded are obvious.²

¹ When a graduate group consisting mainly of experienced supervisors were asked to list observable characteristics indicative of the effectiveness of teaching, most of the items reported were not observable. Considerable study of the problem was necessary before the members of this group could distinguish between observable characteristics and inferred action or evaluations. Monroe, W. S. "Observable Characteristics of Efficiency in Teaching," *Elementary School Journal*, 27: 597-99, April, 1927.

² Reckless, W. C., and Smith, Mapheus. "The Agreement of Three Observers after Practice in Simultaneous Recording of Behavior," *Journal of Applied Psychology*, 18: 635-44, October, 1934. In this study the nature of the observations to be made was obvious from the general instructions. The findings show a rather high degree of agreement.

When the presence of one or more observers is likely to affect the activity of which a record or description is desired, a technique should be devised which will eliminate this difficulty. This is especially important in the observation of young children. Gesell and his associates at Yale University have perfected a one-way vision screen for their study of the behavior of infants.¹

A number of devices have been prepared for the purpose of securing records of the amount and character of pupil participation in classroom activity. Usually these devices are seating charts of the class in which the squares representing the individual pupils are used as the spaces within which to record the symbols showing the extent and character of the pupil participation. Puckett devised symbols to indicate that a pupil asked a question, raised his hand once without being called on, raised his hand another time, was called on and made a good recitation, was called on once when he didn't have his hand raised and made a single word response.² In the device reported by Twitchell³ the symbols are recorded along lines drawn from pupils' names spaced about the chart. Some of the symbols refer to such items as: "called on and responded-verbally or otherwise (without hand raised)," "called on and gave no response or said, 'I don't know,'" "called on (without hand raised) for board work and responded with board work," "volunteered to participate by raising hand and was called on." Horn prepared a device which is somewhat simpler than those of Puckett and Twitchell, and which does not provide for recording so many details.⁴

¹ Gesell, Arnold. *Infancy and Human Growth*. New York: The Macmillan Company, 1928. 418 pp.

The interested reader should consult also

Weiss, A. P. "The Measurement of Infant Behavior," *Psychological Review*, 36: 453-71, November, 1929.

² Puckett, R. C. "Making Supervision Objective," *The School Review*, 36: 209-12, March, 1928.

³ Twitchell, D. F. "An Objective Measure in Supervision," *Journal of Educational Research*, 19: 128-34, February, 1929.

⁴ Horn, Ernest. "Distribution of Opportunity for Participation among the Various Pupils in Class-Room Recitations," *Teachers College, Columbia University, Contributions to Education*, No. 67. New York: Bureau of Publications, Teachers College, Columbia University, 1914, pp. 4-5.

Morrison¹ describes a technique for measuring group attention. The observer takes a position where he can observe easily all of the pupils in a class. The position recommended is the front of the room, but to one side so as to be out of the line of vision between the teacher and the pupils. The observer runs his eyes up and down each row of pupils and at minute intervals records the number of pupils who appear inattentive. The group attention score is obtained by dividing the number of actual minutes of pupil attention by the number of possible minutes of pupil attention. For example, if there are 30 pupils in a class and the class-period is forty-five minutes long, the number of possible minutes of pupil attention is $30 \times 45 = 1350$. If the sum of the numbers of pupils inattentive for all of the minute intervals of the forty-five minute period is 86, then the actual minutes of pupil attention are $1350 - 86 = 1264$. The group attention score is $1264 \div 1350 = .936$ or 94 per cent.

The reader interested in collecting data with respect to individual attention, group attention, or individual-group attention should consult an article by Blume in which he gives an excellent discussion of the techniques and three facsimiles of data sheets used.² The reader will also find it profitable to study the discussions by Knudsen³ and Gray⁴ and the investigations reported by Bjarnason⁵ and Bridges.⁶ In the writings by

¹ Morrison, H. C. *The Practice of Teaching in the Secondary School*. Chicago: The University of Chicago Press, 1926, pp. 119-20.

² Blume, C. E. "Techniques in the Measuring of Pupil Attention," *Scientific Method in Supervision*, The Second Yearbook of the National Conference of Supervisors and Directors of Instruction. New York: Bureau of Publications, Teachers College, Columbia University, 1929, pp. 37-51.

³ Knudsen, C. W. *Evaluation and Improvement of Teaching*. Garden City, New York: Doubleday, Doran and Company, 1932, pp. 263-81.

See also Knudsen, C. W. "A Program of High School Supervision," *Peabody Journal of Education*, 7: 323-32, May, 1930.

⁴ Gray, W. S. "Supervising Instruction in Reading," *Scientific Method in Supervision*, The Second Yearbook of the National Conference of Supervisors and Directors of Instruction. New York: Bureau of Publications, Teachers College, Columbia University, 1929, pp. 181-92. Gray presents a "Group Application and Attention Chart" designed by C. R. Maddox.

⁵ Bjarnason, Loftor. "Relation of Class Size to Control of Attention," *Elementary School Journal*, 26: 36-41, September, 1925.

⁶ Bridges, K. M. B. "The Occupational Interests and Attention of Four-

Blume and by Knudsen evidence is presented to show that when the techniques are used with skill rather highly valid and reliable data are secured with respect to group attention.

When the activity of which a descriptive record is desired extends over a considerable period of time, observation is seldom feasible. In such cases the persons involved may be asked to keep diaries. For example, Flowers¹ sent a request to 170 principals of elementary schools to keep diaries of their daily work for a period of two weeks. Sixty-seven of the principals complied with the request. In using this technique it is well to restrict the record to activities that individuals are capable of observing with respect to themselves. A principal may easily note in a diary the amount of time given to a conference with a parent. He should be able to record, at least so far as the broader aspects are concerned, the nature of the conference. It is doubtful, however, whether the technique will secure dependable data when much introspection, or retrospection, is involved. It is also probable that highly dependable records should not be expected from pupils when they are requested to keep diaries of their activities.

6. *Making estimates and ratings.* Since estimates and ratings are subjective, a scale or score card is usually employed as a means of minimizing the effect of the subjectivity of the process. A scale consists of a series of samples or descriptions of the thing to be rated arranged in order of ascending quality or merit. Usually a numerical value is assigned to each sample. This instrument is illustrated by the scales used for measuring compositions, handwriting, drawings, and the like. A score card consists of a list of the characteristics with respect to which estimates are to be made. Usually the scale of points on which each characteristic is to be rated is given so that when the several ratings are combined they will be appropriately weighted.

The "man-to-man" rating scale, devised in a seminar on "Year-Old Children," *Pedagogical Seminary and Journal of Genetic Psychology*, 36: 551-69, December, 1929.

¹ Flowers, I. V. "The Duties of the Elementary School Principal," *Elementary School Journal*, 27: 414-22, February, 1927.

ducted at the Carnegie Institute of Technology by W. D. Scott, now president of Northwestern University, is a useful device when no better technique is feasible. The rater selects, for example, the best teacher he ever knew and writes the name at the top of a sheet of paper. He next selects the poorest teacher he ever knew and writes his name at the bottom. In between he writes the name of an average teacher, the name of a better-than-average teacher and the name of a poorer-than-average teacher. Numerical ratings such as 15, 12, 9, 6, and 3 may then be assigned to the five names listed. Descriptive statements may be added as a means of assisting the rater to keep in mind the significant characteristics of the teachers of the scale.¹ Freyd² has developed a "graphic" rating scale for rating individuals with respect to twenty traits such as neatness, physique, flexibility, and talkativeness.

When employing a rating scale it is desirable to have several independent ratings made. The average of the several ratings will be more reliable than a single rating. Symonds³ has suggested that in general eight is the minimum number of ratings that should be obtained when the resulting measures are to be considered diagnostic of individuals.⁴

¹ For discussions of the "man-to-man" rating scale, see Rugg, H. O. "Is the Rating of Human Character Practicable?" *Journal of Educational Psychology*, 12: 425-38, 485-501, November, December, 1921; 13: 30-42, 81-93, January, February, 1922.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, pp. 341-43.

² Freyd, Max. "The Graphic Rating Scale," *Journal of Educational Psychology*, 14: 83-102, February, 1923.

For a graphic scale which may be used in rating students with respect to several attitudes, see

Herriott, M. E. "Attitudes as Factors of Scholastic Success," *University of Illinois Bulletin*, Vol. 27, No. 2, *Bureau of Educational Research Bulletin*, No. 47. Urbana: University of Illinois, 1929, pp. 65-72.

³ Symonds, *op. cit.*, p. 354.

⁴ The graduate student or other research worker interested in rating as a means of measuring traits will find the following references helpful:

Brueckner, L. J. "Scales for the Rating of Teaching Skill," *Bulletin of the University of Minnesota*, Vol. 30, No. 12. Minneapolis: University of Minnesota Press, 1927. 28 pp.

Clark, W. W. "Whittier Scale for Grading Juvenile Offenses," *California Bureau of Juvenile Research Bulletin*, No. 11. Whittier, Calif.: Calif. Bureau of Juvenile Research, Whittier State School, Apr., 1922. 8 pp. (*Continued next page*)

Score cards are useful instruments in making estimates since they tend to maintain the same basis of evaluation from one observation to another. A significant limitation lies in the fact that an essential element may be lacking in the thing observed, and yet the total score may be high. For example, in the *Strayer-Engelhardt Score Card for City School Buildings*, 65 points are to be deducted from the perfect score of 1000 points if there are no provisions for fire protection. Hence, a building without fire protection but otherwise well planned and constructed would receive a relatively high score that would tend to be misleading with reference to the actual status of the building for school purposes. While this limitation should be remembered it is not usually a serious one.

In addition to score cards for the rating of school buildings and equipment,¹ and of janitorial and engineering serv-

Cornell, E. L., Coxe, W. W., and Orleans, J. S. *Rating Scale for School Habits*. Yonkers-on-Hudson, New York: World Book Company, 1927.

Haggerty, M. E., Olson, W. C., and Wickman, E. K. *Behavior Rating Schedules, Scales for the Study of Behavior Problems and Problem Tendencies in Children*. Yonkers-on-Hudson, New York: World Book Company, 1930. 11 pp.

Haight, B. F. "A Scheme for Combining Incomplete Rankings," *Journal of Applied Psychology*, 7: 168-72, June, 1923.

Hollingsworth, H. L. *Judging Human Character*. New York: D. Appleton and Company, 1922. 268 pp.

Knight, F. B. "The Effect of the Acquaintance Factor upon Personal Judgments," *Journal of Educational Psychology*, 14: 129-42, March, 1923.

Knight, F. B., and Franzen, R. H. "Pitfalls in Rating Schemes," *Journal of Educational Psychology*, 13: 204-13, April, 1922.

Olson, W. C. *Problem Tendencies in Children; A Method for Their Measurement and Description*. Minneapolis: University of Minnesota Press, 1930. 90 pp.

Shen, Eugene. "The Influence of Friendship upon Personal Ratings," *Journal of Applied Psychology*, 9: 66-68, March, 1925.

Shen, Eugene. "The Reliability Coefficient of Personal Ratings," *Journal of Educational Psychology*, 16: 232-36, April, 1925.

Symonds, P. M. "Notes on Rating," *Journal of Applied Psychology*, 9: 188-95, June, 1925.

Thorndike, E. L. "A Constant Error in Psychological Ratings," *Journal of Applied Psychology*, 4: 25-29, March, 1920.

Wickman, E. K. *Children's Behavior and Teachers' Attitudes*. New York: The Commonwealth Fund, 1928. 247 pp.

¹ The interested reader should consult the following references:

Anderson, C. A. "Tentative Score Card for Elementary School Desks and Seats," *American School Board Journal*, 69: 46-47, July, 1924.

Strayer, G. D. "Score Card for City School Buildings," *Fifteenth Yearbook of*

ice,¹ a number have been prepared for rating such things as health habits, textbooks, teacher characteristics, and supervisory progress. Recently a unique score card has been prepared by Hall for the purpose of rating schools with respect to provisions for gifted children.² The Chapman-Sims Socio-Economic Scale³

the National Society for the Study of Education, Part I. Bloomington, Illinois: Public School Publishing Company, 1916, pp. 41-51.

Strayer, G. D., and Engelhardt, N. L. *Standards for High School Buildings*. New York: Teachers College, Columbia University, 1924. 95 pp.

Strayer, G. D., and Engelhardt, N. L. "Score Card for Village and Rural School Buildings of Four Teachers or Less," *Teachers College Bulletin*, Eleventh Series, No. 9, January 3, 1920. New York: Teachers College, Columbia University, 1920. 22 pp.

Strayer, G. D., Engelhardt, N. L., and Elsbree, W. S. *Standards for the Administration Building of a School System*. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 40 pp.

¹ Engelhardt, N. L., Reeves, C. E., and Womrath, G. F. *Score Card for Public School Janitorial-Engineering Service*. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 6 pp.

² Hall, J. J. "How Does Your School Rate in Providing for Gifted Children?" *Journal of Educational Research*, 22: 81-88, September, 1930.

³ Chapman, J. C., and Sims, V. M. "The Quantitative Measurement of Certain Aspects of Socio-Economic Status," *Journal of Educational Psychology*, 16: 380-90, September, 1925.

Earlier investigations attempting to measure socio-economic status include:

Counts, G. S. "The Selective Character of American Secondary Education," *Supplementary Educational Monographs*, No. 19. Chicago: University of Chicago Press, 1922. 162 pp.

Holley, C. E. "Relationships between Persistence in School and Home Conditions," *Fifteenth Yearbook of the National Society for the Study of Education*, Part II. Chicago: University of Chicago Press, 1916. 119 pp.

Kornhauser, A. W. "The Economic Standing of Parents and the Intelligence of Their Children," *Journal of Educational Psychology*, 9: 159-64, March, 1918.

Van Denburg, J. K. "Causes of the Elimination of Students in Public Secondary Schools of New York City," *Teachers College, Columbia University Contributions to Education*, No. 47. New York: Bureau of Publications, Teachers College, Columbia University, 1911. 206 pp.

The following references will be of interest to the graduate student or other research worker in this connection:

Clark, W. W., and Williams, J. H. "A Guide to the Grading of Neighborhoods," *Publications of Whittier State School*, Department of Research Bulletin, No. 8. Whittier, California: Whittier State School, 1919. 25 pp.

Chapin, F. S. "A Quantitative Scale for Rating the Home and Social Environment of Middle Class Families in an Urban Community: A First Approximation to the Measurement of Socio-Economic Status," *Journal of Educational Psychology*, 19: 99-111, February, 1928.

Moore, E. S. "The Development of Mental Health in a Group of Young Children," *University of Iowa Studies*, Studies in Child Welfare, Vol. IV, No. 6. Iowa City, Iowa, 1931. 128 pp. (A facsimile is given on pp. 96-98 of the Iowa Child Welfare Research Station Scale for Rating of Home Influences.)

(Continued next page)

has been proposed as an instrument for measuring home environment. Heilman used a modification of this scale in his study of "Factors Determining Achievement and Grade Location."¹

7. *Selecting and administering tests.* When adequately defined, the problem specifies the particular abilities or traits to be measured, and these specifications constitute criteria to be used in judging the validity of the test or tests being considered. Determinations of validity have been made for some of the available tests, but a test that is highly valid for one purpose may be unsatisfactory when considered with reference to a different purpose.² When no available test appears satisfactory for securing the data specified by the problem, the investigator should, if he is able to do so, construct his own tests. Curtis³

Terman, L. M., and Goodenough, F. L. "Racial and Social Origin," *Genetic Studies of Genius*, Vol. I. Stanford, California: Stanford University Press, 1925. (On pp. 67-69 is reproduced the Barr Scale for Ratings of Occupational Status.)

Van Alstyne, Dorothy. "The Environment of Three-Year-Old Children," *Teachers College, Columbia University Contributions to Education*, No. 366. New York: Bureau of Publications, Teachers College, Columbia University, 1929. 108 pp. (Describes use of scale devised by Chapin.)

Watson, Goodwin. "A Scale for Rating Home Contributions to Personality Development of Children," *Baltimore Bulletin of Education*, 8: 177-79, May, 1930.

Williams, J. H. "A Guide to the Grading of Homes," *Publications of Whittier State School*, Department of Research Bulletin, No. 7. Whittier, California: Whittier State School, 1918. 21 pp.

¹ Heilman, J. D. "Factors Determining Achievement and Grade Location," *The Pedagogical Seminary and Journal of Genetic Psychology*, 36: 435-57, September, 1929. For a briefer account see "The Relative Influence upon Educational Achievement of Some Hereditary and Environmental Factors," *Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1928, pp. 35-65.

For a facsimile of the revised scale and a discussion of its construction and validation see Heilman, J. D. "A Revision of the Chapman-Sims Socio-Economic Scale," *Journal of Educational Research*, 18: 117-26, September, 1928.

² Errors of validity and errors of measurement are considered in Chapter V. Portions of Chapter VII, which deals with the problem of measurement, may also be read in this connection. The selection of tests in experimental studies is considered in Chapter IX.

³ Curtis, F. D. "Some Values Derived from Extensive Reading of General Science," *Teachers College, Columbia University, Contributions to Education*, No. 163. New York: Bureau of Publications, Teachers College, Columbia University, 1924. 142 pp.

and Irion¹ are examples of research workers who have devised measuring instruments for dealing with their problems.

The investigator should also consider the cost of the test materials and the time required for administration and for scoring the test papers. In some types of studies, the availability of norms and of duplicate forms must also be considered.

The test or tests selected should be carefully administered. If comparisons are to be made with established norms, the published directions for administering must be followed carefully. When such comparisons are not to be made, the directions may be modified, but any changes should be made intelligently, and it is usually important that the directions be the same for all groups of pupils tested.

Application of basic techniques. The preceding exposition of the basic techniques does not adequately reveal the difficulties encountered in collecting data in educational research. In addition to overcoming those inherent in the basic techniques, there are other problems. In experimental studies, an important phase of the process of collecting data is setting up and conducting the experiment.² Sometimes an investigator cannot obtain the data called for by the problem or it is not feasible for him to secure them and he must resort to indirect measurement. For example, certain problems relating to silent reading call for measures of the mental processes involved. A substitute for this information is secured by measuring the eye-movements of the reader. In studies of the supply and demand for teachers, the number of teachers available within a specified area cannot be determined directly.³ Frequently it is not possible, or at least not feasible, to collect data for the entire population specified by the problem. For example, in a study to determine the value of certain measures for predicting teaching success, the problem

¹ Irion, T. W. H. "Comprehension Difficulties of Ninth-Grade Students in the Study of Literature," *Teachers College, Columbia University, Contributions to Education*, No. 189. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 116 pp.

² The procedure of experimentation is described in Chapter IX.

³ For an illustrative study of the supply and demand of teachers see reference 53 at end of Chapter VIII.

may call for measures of teaching success for all students who enter teacher training institutions. Of course, such data can be obtained for only those who actually become teachers and even in the case of this population, measures of teaching success may be difficult or impossible to obtain for all members because the persons involved are scattered over a wide area. When data for only a portion of the population are secured, the dependability of the findings is a function of the representativeness of the sample studied. Hence, the investigator is confronted with the problem of obtaining a sample that is representative or one whose degree of representativeness is known.

When a universe is stratified, that is, consists of relatively homogeneous units or groups, a carefully selected sample from each of the several strata is likely to be highly representative of the total universe. For example, if it is desired to sample pupil achievement in the schools of a large city or of a state, the first step would be to classify the schools on the basis of size or other appropriate criteria. The administration of the test to a few wisely selected classes within each group will be likely to result in a highly representative sample of the total population. When an investigator does not have an opportunity to select a stratified sample, he may be able to present evidence to show that his sample is highly representative. For example, if data have been obtained from the fifth-grade pupils in a given city, he may be able to show that this group of pupils is highly representative of fifth-grade pupils in general by presenting evidence of their general intelligence, chronological age, and other factors which may be expected to influence school achievement.

Occasionally a more mechanical system of sampling may be employed. For example, the students enrolled in a large school may be sampled by taking every fiftieth record from an alphabetical file. The defined words in a dictionary may be sampled by taking the first defined word on the odd numbered pages. The vocabulary of a textbook may be sampled by taking the words in certain lines on every fifth page. Lively and Pressey¹

¹ Lively, B. A., and Pressey, S. L. "A Method for Measuring the 'Vocabulary

have proposed a method of obtaining a thousand-word sample of the vocabulary of textbooks. The procedure of sampling is based upon an estimate of the number of pages which must be sampled, taking one line per page in order to cover one thousand words. For example, a book of five hundred pages with an average of ten words per line would be sampled by taking a specified line on every fifth page throughout the book. Patty and Painter¹ recommend samples that are proportional to the length of the books analyzed.

If the procedure of selection is such that the sample is determined merely by chance, it is called *random*. A random sample is not necessarily representative but any non-representativeness is due to the operation of chance. Frequently in educational research it is difficult to demonstrate that the conditions of random sampling² have been satisfied. They are likely to be approximated when an alphabetical file is sampled by taking records at regular intervals.³ Critical inquiry has revealed that in a number of cases procedures presumed to result in a random sample have failed to do so. For example, if the defined words in a dictionary are sampled by taking the last defined word on each page, Williams⁴ has shown that the sample thus obtained is not random when judged with respect to children's vocabularies. The more commonly used words, whose definitions in general occupy more space than those of unusual words, are more likely to be selected. Hence, a sample obtained by taking the last defined word on pages of a dictionary is biased in favor of the more

Burden' of Textbooks," *Educational Administration and Supervision*, 9: 389-98, October, 1923.

¹ Patty, W. W., and Painter, W. I. "A Technique for Measuring the Vocabulary Burden of Textbooks," *Journal of Educational Research*, 24: 127-34, September, 1931.

² For a statement of these conditions see Yule, G. U. *An Introduction to the Theory of Statistics*, Eighth Edition. London: Charles Griffin and Company, 1927, pp. 259 f.

³ For an illustration of this method of random sampling and evidence of the representativeness of the samples secured, see Wood, Ben D. "The Reliability of Predictions of Proportions on the Basis of Random Sampling," *Journal of Educational Research*, 4: 390-95, December, 1921.

⁴ Williams, H. M. "Some Problems of Sampling in Vocabulary Tests," *Journal of Experimental Education*, 1: 131-33, December, 1932.

commonly used words. Ward¹ has criticized the method of sampling proposed by Lively and Pressey when used as a means of measuring the vocabulary burden of textbooks. Dolch contends that any process of sampling fails to reveal the repetition of words and hence operates to make the book appear more difficult than it really is, because the repetition of words in a text tends to reduce its vocabulary difficulty.² Walker³ describes a number of sampling procedures in which the conditions of random sampling are not satisfied. The sampling that usually occurs when a questionnaire is employed in collecting data is likely not to be random. It appears, therefore, that random sampling is not often possible in educational research. A sample should not be treated as random unless convincing evidence can be cited in support of this characteristic.

A practical question in sampling concerns the size of the sample that may be considered satisfactory. When the sampling is random, the probable error formulae provide a means for determining the number of cases necessary to reduce the probable error of the statistics due to random sampling to any specified limit. In certain types of situations the representativeness of the sample may be estimated by noting the effect of the addition of sub-samples. Suppose a large number of pupil compositions have been accumulated and it is desired to ascertain the types of language errors and their relative frequencies. The investigator may select a series of relatively small random samples and note the changes in his tabulations as additional samples are analyzed. When he finds that no new types of error are being added and the relative frequencies are approximately constant, he is justified in assuming that the total sample is

¹ Ward, J. L. "Measuring 'Vocabulary Burden,'" *American School Board Journal*, 71: 98, September, 1925.

² Dolch, E. W. "Sampling of Reading Matter," *Journal of Educational Research*, 22: 213-15, October, 1930.

³ Walker, H. M., and Students. "The Sampling Problem in Educational Research," *Teachers College Record*, 30: 760-74, May, 1929.

See also Olson, W. C., and Cunningham, E. M. "Time-Sampling Techniques," *Child Development*, 5: 41-58, March, 1934. A bibliography of 76 items is included.

large enough to be highly representative. Johnson and Eurich ¹ have reported a study in which they show that a random sample including 30 per cent of the cases yielded satisfactory results in a certain situation. Although one should not generalize from a single investigation, it is likely that a random sample of this size will usually be satisfactory. In some cases, a smaller random sample may be satisfactory.

This discussion of the application of the basic techniques suggests only a few of the difficulties that investigators actually encounter in collecting data. The particular difficulties are so varied that a complete treatment is not feasible in this volume. A person, especially one inexperienced in educational research, who contemplates the study of a problem, should consult reports of similar researches with respect to the techniques employed in collecting data. The bibliography of survey investigations at the end of Chapter VIII and that of experimental studies at the end of Chapter IX are recommended to readers who are interested in extending their study of collecting data in educational research.

¹ Johnson, D. A., and Eurich, A. C. "An Empirical Test of Sampling," *Journal of Experimental Education*, 3: 174-79, March, 1935.

CHAPTER IV

ELEMENTARY TECHNIQUES FOR HANDLING DATA ¹

A. SCALES AND CALCULATIONS

The meaning of numerical data. Numerical data consisting of counts of things such as the members of a class, full-time teachers employed, books in libraries, and the like are to be regarded as *exact*. The scale of measurement on which such measures are expressed is called discontinuous or *discrete* to designate that it has values at only the points designated by integers. Test scores and most other types of data dealt with in educational research are expressed on a *continuous* scale of measurement. Given sufficiently precise measuring instruments, there is no reason why test scores and other such measures may not be expressed to any desired number of decimal places. In practice, however, such data are usually expressed as integers. In most cases the integer employed is the one marking the lower limit of the unit division of the scale in which the exact value of the measure lies. For example, a test score of 27 means that the exact measure is not less than 27.0 and is not as much as 28.0. In other words, although the score is given as 27, the exact value may be anywhere between 27.0 and 28.0. Occasionally, measures are expressed in terms of the nearest division point of the scale. For example, a measure of the quality of handwriting expressed as 8 means that the exact measure of the quality may be anywhere between 7.5 and 8.5. In both cases the data are to be regarded as only *approximate*.

¹ Other statistical techniques are described in later chapters in connection with the consideration of types of problems to which they are applicable. The reader may locate these techniques by employing the topical index at the end of the volume.

The scale of measurement on which teachers' salaries are expressed is theoretically continuous, but in practice it is discontinuous. Salaries are usually round numbers such as \$900, \$1000, \$1080, \$1850, and the like. A teacher is seldom, if ever, paid a salary of \$1203.42. Hence, teachers' salaries are to be regarded as exact measures, although the scale of measurement is theoretically continuous.

The number of decimal places or significant figures in the results of calculation. The fact that much of the data with which we deal in educational research are only approximate raises the question of the number of decimal places or significant figures to be retained in the results of calculation. The situation is complicated by the presence of errors and other data faults which are to be described in Chapter V, but it will be helpful to consider the question merely on the basis of the nature of approximate measures. In the case of the addition or subtraction of approximate numbers the sum or difference should include only as many decimal places as appear in the least precise of the items. For example, if 3.9, 8.475, and 1.2846 are to be added, the addends should be rounded off to one decimal place or the sum should be rounded off to one decimal place. Writing the sum with a larger number of decimal places will tend to give a false impression of its precision. If the mean is calculated by dividing the sum by the number of items, it may be expressed to one additional decimal place. If the number of items is relatively large, say 50 or more, a second additional decimal place is appropriate. When a mean is calculated from a frequency distribution, more conservative rules should be followed.

In the case of multiplication or division, we are concerned with the number of significant figures rather than the number of decimal places. Zeros employed merely to locate the decimal points are not considered significant. For example, each of the measures 742, 7.42, .0742, and .000742 is described as having three significant figures. A zero between two other figures is counted in determining the number of significant figures. A

zero introduced in rounding off a number is not counted as significant. For example, if the population of the city is given as 12,500, meaning thereby that the exact number of inhabitants is nearer 12,500 than 12,400 or 12,600, the number of significant figures is three. If, however, the population is known to be exactly 12,500, the number of significant figures is five. As a means of indicating that an integral number ending in one or more zeros is to be considered exact, a decimal point may be placed after the last zero. Thus 12,500. would be understood as an exact number.

In multiplication or division of approximate numbers, the general rule is that the number of significant figures in a product or quotient should not exceed that of the measure having the smaller number of significant figures. For example, if the diameter of a cross-section of a tree has been measured as 14 inches and the circumference is calculated by multiplying by 3.1416, the result should be given as 44 and not as 43.9824. The items should not be rounded off before the multiplication or division is performed and in the case of division the quotient should be calculated to two places more than are to be retained.

A square root may be carried to as many significant figures as there are in the number. The standard deviation ¹ may be calculated to one or two decimal places beyond the number appearing in the measures. For example, if the measures are in terms of integers, the calculation of the standard deviation should be carried to one or two decimal places. In accomplishing this, it is advisable to carry the calculations under the radical to four places.

The reader should bear in mind these rules are derived from the nature of approximate measures. If data faults, which are dealt with in Chapter V, are also considered, more conservative rules would be indicated. Since data faults vary, it is not possible to state specific rules, but the investigator should restrict the number of decimal places or significant figures so that the precision of the calculated statistic will not be grossly misrepre-

¹ See page 75.

sented. In some cases one may be guided by prevailing practice. Coefficients of correlation are usually expressed to two decimal places and conformity with this practice is appropriate. When the results of calculations are listed in a table, items of the same type should have the same number of decimal places.

Computational aids. The arithmetical calculation involved in handling numerical data frequently requires much time. The research worker may, if his resources permit, employ clerical assistance for this routine work, but, in order to insure accuracy, the calculations must be checked. This checking requires additional time. Hence, it is desirable to utilize such aids as may be available.

Tables of products ¹ are useful for securing products and quotients. Tables of logarithms, familiar to one who has studied trigonometry, are useful in multiplying, dividing, raising to a power, and extracting roots.² Tables of squares ³ are useful in the calculation of the coefficient of correlation. In calculating probable errors and coefficients of partial correlation, tables giving values of $1 - r^2$ and $\sqrt{1 - r^2}$ facilitate computation.⁴ The tables prepared by Pearson ⁵ may be recommended for advanced statistical work. The handbook by Dunlap and Kurtz ⁶ is a very useful tool. It contains several tables, includ-

¹ A. L. Crelle's *Calculating Tables*. Berlin: Walter de Gruyter and Co., 1919. (New edition by O. Seeliger.) These tables give all products up to 999 times 999.

Peters, J. *Neue Rechentafeln für Multiplikation und Division*. Berlin: Druck und Verlag G. Reimer, 1909. These tables give all products up to 99 times 99,999.

² For an explanation of how to use logarithms, see Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, pp. 47-64.

³ For example, Barlow's *Tables of Squares, Cubes, Square Roots, Reciprocals of All Integer Numbers up to 10,000*. New York: Spon and Chamberlain, 1927. 200 pp.

⁴ Holzinger, K. J. *Statistical Tables for Students in Education and Psychology*. Chicago: University of Chicago Press, 1925.

Miner, J. R. *Tables of $\sqrt{1 - r^2}$ and $1 - r^2$ for Use in Partial Correlation and in Trigonometry*. Baltimore: The Johns Hopkins Press, 1922. 49 pp.

⁵ Pearson, Karl. *Tables for Statisticians and Biometricians*. Cambridge: Cambridge University Press, 1914. (Second edition, 1924; First edition, Vol. 2, 1931.) 143 pp.

⁶ Dunlap, J. W., and Kurtz, A. K. *Handbook of Statistical Nomographs, Tables, and Formulas*. Yonkers-on-Hudson, New York: World Book Company, 1932. 163 pp.

ing squares, square roots, reciprocals and logarithms of numbers; and a number of nomographs ¹ useful in the calculation of certain statistics, such as standard error of a mean, intelligence quotient, and bi-serial coefficient of correlation. It also lists practically all of the formulae used in the statistical treatment of educational data.

Arithmetical calculations may be accomplished by means of machines. The Monroe and the Marchant are perhaps the most useful. Although practice is required for skillful use of a calculating machine, the manipulations for the several arithmetical operations are easy to learn and the procedure for calculating a given statistic will soon be memorized. When an appropriate machine is available, much time can be saved by using it. In some cases the maximum economy of time requires a formula adapted to the machine. See page 94 for an illustration. The Hollerith Machine is useful in tabulation and classification when the data are extensive. It may also be used in calculating coefficients of correlation ² and in solving regression equations.³ A slide rule may be used when a high degree of accuracy is not required but its principal use is in checking calculations accomplished by other means.

B. STATISTICS OF FREQUENCY DISTRIBUTIONS

Summarizing a group of numerical data. When an investigator has a collection of data such as test scores, teachers' salaries, number of volumes in certain school libraries, and the like, he usually calculates the arithmetical mean (average) or median

¹ A nomograph is a graphical device, somewhat similar in principle to the slide rule, from which the results of a series of calculations can be read. The readings are only moderately precise. The construction of a nomograph has been described by Griffin in the following reference. It includes an annotated bibliography of 73 references.

Griffin, Harold D. "How to Construct a Nomogram," *Journal of Educational Psychology*, 23: 561-77, November, 1932.

² Mendenhall, R. M., and Warren, R. "Computing Statistical Coefficients from Punched Cards," *Journal of Educational Psychology*, 21: 53-62, January, 1930.

³ Segel, David. "The Automatic Prediction of Scholastic Success by Using the Multiple Regression Technique with Electric Tabulating and Accounting Machines," *Journal of Educational Psychology*, 22: 139-44, February, 1931.

as a central tendency or summary measure.¹ The standard deviation or other measure of variability furnishes a further description of the group of data. These measures may be calculated directly from the data. For example, if X is used to represent the items of data, N the number of items, and Σ (capital sigma) the summation or addition of the items, the procedure for calculating the mean may be expressed by M (mean) = $\frac{\Sigma X}{N}$.

(See page 76 for calculation of the standard deviation.) Usually, however, it is desirable to tabulate the measures in a frequency distribution. When this is done, the mean, standard deviation, and similar measures are usually calculated from the frequency distribution. This procedure involves an assumption that will be noted later. Unless this assumption is approximated, the result obtained will be likely to differ from that secured by direct calculation from the items.

Constructing a frequency distribution. A frequency distribution consists of a scale of intervals and of the number of measures or items of data falling in each interval. The first step in constructing a frequency distribution is that of dividing the scale from zero or from a convenient point below the lowest measure to a point at or slightly above the highest measure into a convenient number of equal intervals or steps.² There is no standard number of intervals, but usually an effort is made to have not fewer than ten nor more than twenty-five. The choice of the end-points of the intervals should be guided by three principles. In the first place they should be consistent with the nature of the data. If the measures are approximate and expressed in terms of lower limits of the unit divisions of the scale of measurement, the end-points should be at unit

¹ The mode (see page 74) is sometimes calculated. Certain situations require the use of the geometric mean defined by the formula $\log (GM) = \frac{1}{N} \Sigma \log X$ or the harmonic mean defined by $\frac{1}{H} = \frac{1}{N} \Sigma \left(\frac{1}{X} \right)$.

² Occasionally frequency distributions are formed with unequal intervals, especially at the extremes. This, however, is seldom done unless the data are of such a nature as to make it desirable.

division points of the scale. For example, if 17 means at least 17, but not as much as 18, the precise limits of the interval 15 to 19 are 15.00 and 19.99. . . . If the measures are expressed in terms of the nearest integer, for example, if 17 means a value falling between 16.50 and 17.49 . . . , the lower limit of the interval is still written 15 and the upper one 19. The precise limits, however, are 14.50 and 19.49. . . .¹ The second principle is that the choice of the end-points should facilitate tabulation. End-points that are multiples of 5, 10, or 100 are likely to be more convenient than ones with other endings. Finally, the end-points should be chosen so that the average values of the measures in the intervals will approach as nearly as possible the mid-points of the respective intervals.

The intervals may be designated by writing their lower limits as in Table I, but, if desired, the upper limits may also be designated. The tabulation sheet is prepared by writing the scale intervals at the left side preferably using ruled paper. If the measures to be tabulated are on separate sheets or cards, they may be sorted according to scale intervals and the number of measures in each pile counted and recorded in the frequency column of the distribution. If sorting is not feasible or is not considered desirable, the measures may be tallied as shown in Table I. In tallying, each group of five is represented by four downward strokes, ||||, and a fifth diagonal stroke, \. This device aids in counting up the frequencies. The final step is to write the frequencies opposite their respective scale intervals. Generally it is desirable to copy the scale intervals and their frequencies on another sheet. The frequency distribution should be given an appropriate caption as a means of labeling it.

For some purposes a cumulative frequency distribution is useful. It is formed by totaling the frequencies for the successive intervals. Thus, in Table I the cumulative frequencies would be 1, 4, 8, 16, 20, etc.

Graphical representation of frequency distributions. Many persons are aided in comprehending a frequency distribution by

¹ See footnote on page 69.

having it represented graphically. The scale of the distribution is laid off on a horizontal line and the frequencies are represented by vertical distances. Graphical representation is treated briefly in Chapter VIII. For further details the reader should consult a text that deals with this topic.

Describing a frequency distribution. A frequency distribution is described by specifying its shape and determining its central tendency and a measure of its variability or spread. The frequency distributions of large unselected groups of educational data approach the normal shape and unless a statement to the contrary is made, an approximation to a normal distribution is understood. The mean or the median is usually calculated as the central tendency. The standard deviation and the median deviation (probable error) are the most frequently used measures of variability.

TABLE I. SHOWING THE TABULATION OF A FREQUENCY DISTRIBUTION

SCALE INTERVALS	TALLIES	FREQUENCIES
475		1
450		1
425		1
400		0
375		0
350		0
325		3
300		3
275		3
250		3
225		4
200		4
175		4
150		4
125		4
100		4
75		4
50		4
25		4
0		4
$N =$		81

Calculating the mean of a frequency distribution. The first step is to assume a mean. This may be chosen at any point,

but the calculation will be reduced to a minimum if the assumed mean is chosen at the mid-point of the interval in which the true mean falls. In Table II the mean has been assumed to fall at the mid-point of the interval from 200 to, but not including, 225, or at 212.5.¹ The eight scores in the interval 225 to, but not including, 250, are assumed to be uniformly distributed over it and hence "on the average" are 25 units, or one scale interval, above the assumed mean. In order to reduce the calculations to a minimum, this deviation is called one interval. In the same way the deviations of the other scores from the assumed mean are expressed in terms of intervals. A negative deviation means that the scores fall in an interval below the assumed mean. The deviations are given in the third column of the table. In the fourth column of the table the products of the deviations and frequencies are recorded. The sum of the positive products is 80, and the sum of the negative ones is -132. The difference is -52. This is divided by the total of the frequencies, which gives a quotient of $-.642$ of an interval. Since the interval is 25 units, this quotient is multiplied by 25 to find the correction in terms of units. This correction added *algebraically* to the assumed mean gives the true mean of 196.45.²

By employing symbols³ the procedure just described may

¹ The mid-point of a scale or class interval depends on the meaning of the units employed. When a measure such as a score on a spelling test means, for example, 16 up to but not including 17, the mid-point of the interval 15-20 (more precisely 15.00-19.99 . . .) is 17.5. If, however, a score of 16 means exactly 16, or means 15.50 to 16.49 . . ., then the interval 15-20 (more precisely 14.50-19.49 . . .) has as its mid-point 17. Most measures of achievement and intelligence are interpreted in the same fashion as spelling scores. Quality of handwriting, merit of compositions, and weights of children are usually expressed in terms of the nearest integer and the second procedure should be employed.

² In order to avoid giving a false impression of precision this result should be recorded as 196.5. See page 62.

³ A group of measures is commonly represented by the symbol X . When there are two or more groups of measures subscripts are usually employed but occasionally Y is used to designate a second group. When the measures are expressed as deviations from the mean of the group, small letters are used as symbols. For further exposition of statistical symbolism and a list of symbols see Appendix.

TABLE II. ILLUSTRATING THE CALCULATION OF THE MEAN FROM A FREQUENCY DISTRIBUTION

SCALE INTERVALS	FREQUENCY f	DEVIATION IN INTERVALS x'	fx'
475	1	11	11
450	1	10	10
425	1	9	9
400		8	
375		7	
350		6	
325	3	5	15
300	3	4	12
275	3	3	9
250	3	2	6
225	8	1	8
200	10	0	80
175	13	-1	-13
150	15	-2	-30
125	4	-3	-12
100	8	-4	-32
75	4	-5	-20
50	3	-6	-18
25	1	-7	-7
0			-132
N	81		

Assumed mean..... 212.50

Correction..... -16.05

True mean..... 196.45

$$\text{Correction} = i \frac{\sum fx'}{N}$$

$$= 25 \frac{80 - 132}{81} = 25(-.642)$$

$$= -16.05$$

be designated by means of a simple formula. Let M represent the mean to be calculated, M' the assumed mean, x' the deviations of the intervals from the assumed mean,¹ f the frequency of the measures in an interval, and i the width of an interval.

¹ The symbol d' is sometimes used for this purpose.

Then

$$M = M' + i \frac{\Sigma fx'}{N}$$

In this formula Σ (capital sigma) indicates the sum of the various products formed by multiplying the frequency of an interval (f) by its deviation (x'). The number of cases or measures is designated by N .

This method of calculating the mean from a frequency distribution assumes that the average value of the measures within an interval falls at the mid-point of the interval. When this assumption is not satisfied or departures from it are not neutralized, the calculated mean will be affected. The assumption is not fully satisfied in a normal distribution if the intervals are large. The means of the measures of the several intervals will be slightly nearer the center of the distribution than the corresponding mid-points, but the effects will be neutralized. When the distribution is not normal, non-conformity with the assumption is likely to affect the results of the calculation. If the non-conformity is marked, the error may be distinctly significant, especially when the grouping is coarse. For example, teachers' salaries are usually concentrated at certain points. If the points of concentration are not near the middle of the intervals chosen, the calculated mean may be significantly in error. It has been suggested that when N is less than 50 and the width of the interval is greater than one, calculation should be made from the data.

It should be noted that when the measures are expressed in terms of the lower limits of the unit divisions of the scale of measurement, the mean calculated by dividing the sum of the items by their number will be too small. When the mean is calculated from a frequency distribution, the fact that the measures are expressed in terms of their lower limits is taken into account and a more accurate value will usually be obtained. The mean calculated directly from the scores tabulated in Table II is 196.04. The difference between this value and the

mean calculated from the frequency distribution is probably typical of the difference to be expected in such cases. If the measures are expressed in terms of the nearest division point of the scale of measurement, the calculation of the mean by adding the measures and dividing the sum by the number of items will give an accurate result. The mean of such measures calculated from a frequency distribution will be slightly too large unless the limits of the intervals are properly chosen or the nature of the measures is recognized in the value assigned to the assumed mean.

Occasionally the research worker is confronted with the necessity of calculating the mean from a frequency distribution whose intervals are not equal in width. When this condition prevails, it is only necessary that the deviations of the mid-points of the several intervals from the assumed mean be correctly expressed.

Calculating the median. As the term implies, the median is the *point on the scale* on each side of which one-half of the measures fall.¹ Sometimes this term is used in the sense of *mid-measure*, which is defined as the middlemost measure of a series of measures arranged in ascending or descending order of magnitude. If the number in the series is odd, the mid-measure is that one above and below which there is an equal number of measures. If the number of measures is even, the mid-measure is taken as the average of the two mid-most items. If the scale of measurement is continuous the meaning of the numerical expressions of the measures should be recognized. For example, if the mid-measure is the third of five expressed as 17, its value should be 17.5 rather than 17.

The median of a frequency distribution is calculated by means of the following formulae which serve as a check on each other:

$$Md = l + \frac{\left(\frac{N}{2} - S_l\right)}{f} i$$

¹ If considered critically in the case of a limited number of measures, this definition is indefinite. For practical purposes, the most precise definition is expressed by the formulae for calculating the median.

$$Md = u - \frac{\left(\frac{N}{2} - S_u\right)}{f} i$$

The symbol l refers to the lower limit of the class interval in which the median is estimated to fall; u is the symbol for the upper limit of this class interval; S_l designates the sum of the measures *below* the lower limit of the class interval in which the median is estimated to fall; and S_u , the sum of the measures *above* the upper limit of this interval. The symbol i refers to the width of the scale interval and the symbol f designates the number of measures in the class interval within which the median is estimated to fall. The use of these formulae is illustrated by the following calculation of the median of the distribution given in Table II.

$$Md = 175 + \frac{\frac{81}{2} - 35}{13} \times 25 = 185.58$$

$$Md = 200 - \frac{\frac{81}{2} - 33}{13} \times 25 = 185.58$$

The choice of the scale interval 175 up to, but not including 200, is made as the result of inspection. Approximately 40 measures must be above and below the median. Hence, counting in from either end soon reveals that the median is in the interval chosen.

Other points in frequency distributions. Other points on the scale of a frequency distribution may be obtained with slight adaptation of the formulae given above. Q_1 , the first quartile, the point above which are three-fourths of the measures and below which are one-fourth of the measures, can be obtained by inserting, respectively, $\frac{N}{4}$ and $\frac{3N}{4}$ for $\frac{N}{2}$ in the formulae and modifying the meaning of l and u accordingly. Q_3 , the third quartile, the point below which are three-fourths of the measures

and above which are one-fourth of the measures, can be obtained by inserting, respectively, $\frac{3N}{4}$ and $\frac{N}{4}$. The decile points, dividing the scale in tenths, may be obtained by substituting in the first formula $\frac{N}{10}, \frac{2N}{10}, \frac{3N}{10}, \frac{4N}{10}, \frac{5N}{10}, \frac{6N}{10}, \frac{7N}{10}, \frac{8N}{10}$, and $\frac{9N}{10}$. These computations may be checked by substituting for $\frac{N}{2}$ in the second formula $\frac{9N}{10}, \frac{8N}{10}, \frac{7N}{10}, \frac{6N}{10}, \frac{5N}{10}, \frac{4N}{10}, \frac{3N}{10}, \frac{2N}{10}, \frac{N}{10}$. Percentile points may be secured by similar substitution of appropriate fractions. For example, the 63 percentile point would be obtained by substituting in the first formula $\frac{63N}{100}$ and in the second $\frac{37N}{100}$.

The mode. The mode is occasionally used in describing a frequency distribution. In a simple series of measures, the mode is the most frequently occurring measure. In a frequency distribution, the crude mode is defined as the scale interval which has the largest frequency. This measure of central tendency, however, is obviously unsatisfactory because it depends upon the grouping of the measures made in forming the frequency distribution. Several formulae have been proposed for calculating a modal point, but they are seldom used.¹

Calculating measures of variability of a frequency distribution. A mean or median represents only the central tendency of a frequency distribution. Since such assemblages of data differ in the spread or variability, a measure of this characteristic is a useful supplementary description. The *range* is an easily understood measure of variability. It is the distance on the scale of measurement from the lowest to the highest measure. A single extreme case may influence the size of this measure unduly. Hence, the 10 – 90 *percentile range*, D_{10-90} , is fre-

¹ Five formulae for calculating the mode are given by Dunlap, J. W., and Kurtz, A. K. *Handbook of Statistical Nomographs, Tables, and Formulas*. Yonkers-on-Hudson, New York: World Book Company, 1932, p. 105.

quently used in place of the total range. It is obtained by subtracting the value of the tenth percentile point from that of the ninetieth percentile point. The *quartile deviation*, Q , is obtained by subtracting Q_1 from Q_3 and dividing the difference by 2. When this distance is measured off on both sides of the median, the space thus defined includes 50 per cent of the measures, provided the distribution is normal in shape. Another measure of variability is the average deviation (AD). As the term suggests, it is the mean of the deviations from the median or mean, the sign of the deviations being disregarded. It, however, is not very frequently used.¹

The most commonly used measure of variability is the *standard deviation* which is the square root of the mean of the squares of the deviations from the mean.² This statistic is generally represented by the symbol ³ σ . When the measures are not grouped in a frequency distribution, the calculation is indicated by the formula

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

For a frequency distribution, it is defined by the formula

$$\sigma = i \sqrt{\frac{\sum fx^2}{N}}$$

In this formula Σ (capital sigma) designates the summation of all products of f and x^2 , and i , the width of the intervals or steps of the frequency distribution.

Table III illustrates the calculation of the standard deviation from the frequency distribution. The procedure, up to a cer-

¹ For the procedure to be used in computing the average deviation see any standard statistical text.

² This is the original definition of the standard deviation and the one generally recognized. Unfortunately, some writers have defined it in other terms. For a discussion of the various definitions, see Eells, W. C. "A Plea for a Standard Definition of the Standard Deviation," *Journal of Educational Research*, 13: 45-52, January, 1926.

³ This symbol is the small Greek letter "sigma." Occasionally SD is used to designate standard deviation, but this practice is not recommended.

TABLE III. ILLUSTRATING THE CALCULATION OF THE STANDARD DEVIATION OF A FREQUENCY DISTRIBUTION

SCALE INTERVALS	<i>f</i>	<i>x'</i>	<i>fx'</i>	<i>f(x')²</i>
475	1	11	11	121
450	1	10	10	100
425	1	9	9	81
400		8		
375		7		
350		6		
325	3	5	15	75
300	3	4	12	48
275	3	3	9	27
250	3	2	6	12
225	8	1	8	8
200	10	0	80	
175	13	-1	-13	13
150	15	-2	-30	60
125	4	-3	-12	36
100	8	-4	-32	128
75	4	-5	-20	100
50	3	-6	-18	108
25	1	-7	-7	49
0				
<i>N</i>	81		-132	966

$$\begin{aligned}
 \sigma &= 25\sqrt{\frac{966}{81} - .642^2} \\
 &= 25\sqrt{11.5138} \\
 &= 25(3.39) \\
 &= 84.75
 \end{aligned}$$

tain point, is identical to that for calculating the mean.¹ In the last column of the table, the products of the frequency and the square of the deviation of each interval are entered. The deviations have been expressed from the assumed mean 212.5. The true mean is 196.45. It is, therefore, necessary to correct for the error introduced by this assumption. This error is $-.642$ of an interval. The procedure for making the correction is indicated by the following formula in which c designates the difference between the true and the assumed means.

¹ See page 70. The deviations are taken from an assumed mean.

$$\sigma = i\sqrt{\frac{\sum f(x')^2}{N} - c^2}$$

In terms of raw, ungrouped measures, the calculation of the standard deviation may be accomplished as indicated by the formula

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum X^2}{N} - \frac{(\sum X)^2}{N^2}} \\ &= \sqrt{\frac{\sum X^2}{N} - M^2}\end{aligned}$$

This procedure avoids the necessity of setting up the frequency distribution and is economical when an appropriate calculating machine is available. For example, using a Monroe Calculating Machine, lock the "1" of the extreme left column of the keyboard and shift the carriage to the left. Then punch, successively, at the right of the keyboard, the various values of X , multiplying each by itself. The final readings on the lower dial will be $\sum X$ at the left and $\sum X^2$ at the right. The mean is found by dividing $\sum X$ by N . The subtraction of M^2 may be accomplished by subtracting the mean M times from the quotient obtained by dividing $\sum X^2$ by N .

The median deviation, MdD , is the median of the deviations from the mean. The term *probable error*, PE , has the same meaning but should be used only when the distribution is one of errors. The median deviation of a distribution is customarily computed by multiplying the standard deviation (standard error) by the constant .6745.

$$MdD = .6745\sigma$$

This procedure assumes a normal distribution or one which deviates from it very slightly.

Additional items of description when the distribution is not normal. When a frequency distribution does not approach the normal curve in shape, the central tendency and a measure of variability do not furnish an adequate description, especially

when the departure from the normal shape is marked. If the measures on one side of the mean or median tend to be bunched and there is tailing out on the other, the distribution is said to be *skewed*. In such cases, it is desirable to obtain a measure of this abnormality or skewness. In a normal distribution the mean and the median coincide. In a skewed distribution they do not, and the extent to which these measures are separated affords an index of the degree of skewness. The following formula is one commonly used.¹

$$Sk = 3 \frac{(M - Md)}{\sigma}$$

This formula indicates, in addition to the magnitude of the skewness, whether it is "positive" or "negative" in character. A positively skewed curve is one which is drawn out further to the right than to the left, the mean is greater than the median in scale value, and the curve slopes steeply on the left, but gently on the right. A negatively skewed curve is one in which the curve is drawn out more to the left, the median is greater than the mean, and the slope on the left is gentle but steep on the right.

A distribution may tend to be symmetrical about the mean and hence approximate zero in skewness but exhibit other abnormalities. If it tends to be rectangular in shape, the standard deviation will tend to be a misleading measure. In such a case, it is desirable to calculate the kurtosis which is defined by the formula:²

$$Ku = \frac{N \sum x^4}{(\sum x^2)^2} - 3$$

For a normal distribution, the kurtosis is zero.

Measures of skewness and of kurtosis are not often calculated,

¹ Dunlap, J. W., and Kurtz, A. K. *Handbook of Statistical Nomographs, Tables, and Formulas*. Yonkers-on-Hudson: World Book Company, 1932, p. 112. Eight formulae are given.

² This formula is written in other forms. A common form is $Ku = \frac{\mu_4}{\mu_2^2} - 3$. Sometimes the "3" is omitted.

but unless the shape of a distribution is approximately normal, a description limited to a central tendency and a measure of variability is not complete.

Identification of normal distributions. The normal curve, which is approximated in shape by the graphical representation of the frequency distribution of unselected groups of many types of educational data, is defined by the equation

$$y = \frac{N}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

in which N is the total of the frequencies of the distribution and e the base of the Napierian logarithms. A normal distribution is also obtained by expanding the binomial $(p + q)^n$ and taking the coefficients of the terms of the resulting polynomial as the frequencies. By making n large, the graphical representation of these frequencies will approach a smooth curve.

A measure of skewness affords an indication of the degree of departure from the normal shape, but it should be regarded as only a crude index of the degree of normality. A distribution may be symmetrical, the opposite of skewed, and not be normal. A more precise means of discovering whether or not a given distribution departs significantly from the normal shape is Pearson's Chi-Square Test.¹ The use of this test demonstrates whether or not the normal curve fits the observed data within the fluctuations of random sampling.

¹ Pearson, Karl. "On the Probability that Two Independent Distributions of Frequency Are Really Samples from the Same Population," *Biometrika*, 8: 250-54, 1911.

Pearson, Karl. "On a Brief Proof of the Fundamental Formula for Testing the Goodness of Fit of Frequency Distributions and on the Probable Error of P ," *Philosophical Magazine*, 30: 369, 1916.

Pearson, Karl. "On the χ^2 Test of Goodness of Fit," *Biometrika*, 14: 186-91, 1922.

Pearson, Karl. "Further Note on the χ^2 Test of Goodness of Fit," *Biometrika*, 14: 418, 1922.

Explanation and illustration of the use of this test are given in

Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, pp. 245-48.

The relation between deviations from the mean and corresponding areas under the normal curve. The normal curve has a number of interesting properties,¹ but in educational research we are concerned mainly with the fraction of the total area that is cut off when perpendiculars are erected at points on the base line. The area cut off in this way may be determined by the methods of integral calculus,² but the calculation is laborious and tables have been prepared from which the areas corresponding to various deviations from the mean may be read. These tables vary in form and in the labels employed to designate the quantities given. The abscissa distances (deviations from the mean) are usually given in terms of σ as a unit and this fact is indicated by labeling the column $\frac{x}{\sigma}$. A few authors express the abscissa distances in terms of PE . The area given is usually that included between the mean ($x = 0$) and the designated value of $\frac{x}{\sigma}$. Several symbols have been used for designating this area. In his text³ Holzinger, following the lead of Sheppard and Pearson, uses $\frac{1}{2}\alpha$ but in his volume of tables⁴ he employs the expression "area from $\frac{x}{\sigma} = 0$ to given $\frac{x}{\sigma}$." In the Kelley-Wood Table⁵ I is used as the symbol. Sometimes the area given is that from the extreme left up to the point defined by $\frac{x}{\sigma}$. This area is designated by $\frac{1}{2}(1 + \alpha)$ by Sheppard and Pearson. Kelley uses p . The total area is usually taken as 1.000 but Garrett⁶ uses 10,000. The choice of the total area is immaterial when only the fraction of the total area is desired.

¹ For an account see Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, Chapter V.

² See Rietz, H. L., et al. *Handbook of Mathematical Statistics*. Boston: Houghton Mifflin Company, 1924, p. 14 for the probability integral.

³ *Statistical Methods for Students in Education*, p. 211.

⁴ Holzinger, K. J. *Statistical Tables for Students in Education and Psychology*. Chicago: University of Chicago Press, 1925, Tables XI and XII.

⁵ Kelley, *op. cit.*, Appendix C.

⁶ Garrett, H. E. *Statistics in Psychology and Education*. New York: Longmans Green and Company, 1926, p. 91.

The ratio of a section of the area to the total area under the normal curve expresses the probability that a datum (measure) selected at random from the population represented by the distribution is within the section specified. For example, if the section is that marked off by measuring $.6745\sigma$ in both directions from the mean, the ratio of this area to the total area is $\frac{1}{2}$. Hence, the chances are one to one that a datum selected at random will fall within this section. If the section specified is that to the left of the point $M - .6745\sigma$, the corresponding ratio is $\frac{1}{4}$. The chances are one to three, or one out of four, that a datum selected at random will fall within this section. Given a normal distribution, a measure of its variability (standard deviation or median deviation) and the mean, the probability that a datum selected at random from the population will fall within certain limits can be determined from the specifications of these limits. These limits may be between two points defined by measuring specified distances from the mean or they may be designated as merely beyond, i.e., to the right or to the left of a specified point. If the probability is given, the limits corresponding to it may be determined.

Most of our use of a table of the values of the probability integral is with three distributions: (1) that formed by the values of a statistic computed from a large number of similar but independent random samples from a universe or large population, (2) that of the variable errors included in data, and (3) that of the errors of estimate when predictions have been made within a typical population. The first distribution may also be thought of as the distribution of the differences between the calculated values of the statistic and the value of the statistic for the universe or larger population from which the samples have been taken. These differences may be regarded as the errors of the calculated values of the statistic when they are taken as the value of the statistic for the universe or larger population. Hence, the distribution may be referred to as one of errors. The standard deviation of a distribution of one of

these types is frequently designated the *standard error*, and the median deviation is properly called the *probable error*.

None of these distributions is actually formed in typical statistical work, but assuming that they are normal, it is possible under certain conditions to calculate their standard deviations. Hence, either of the two determinations noted above may be easily accomplished with the aid of a table of the values of the probability integral. The distribution of the values of a statistic from similar but independent random samples or the distribution of the corresponding errors will be referred to later in the present chapter. The distribution of variable errors in data is dealt with under the head of the probable error of measurement in Chapter V. Errors of estimate are considered in Chapter X.

C. CALCULATION OF COMPARABLE MEASURES

Calculating comparable measures. Two sets of measures considered to describe magnitudes of the same thing¹ are comparable provided zero on one scale of measures is equivalent to zero on the second and the units of the two scales are such that n units on one may be considered to designate a magnitude equivalent to that represented by n units on the other. When these conditions do not prevail, a prerequisite for comparisons is the reduction of one set of measures to the basis of the other or the reduction of both to a common basis.

The procedures most commonly employed for effecting a reduction are based on the assumptions that the means of two or more groups of measures of the same thing, or certain points measured from them, qualify as a common zero point, that the standard deviations of the distributions may be regarded as designating equivalent magnitudes, and that the ratio of corresponding deviations from the mean is the same as that given

¹ The meaning of "same thing" should be noted. From one point of view a silent reading test and an arithmetic test do not measure the same thing. One measures what we call silent reading ability, and the other arithmetical ability. Both, however, measure achievement and from this point of view scores yielded by them may be considered to be measures of the same thing. Similarly, measures of height and measures of weight may be considered measures of physical maturity.

by the standard deviations.¹ Employing symbols, these assumptions are expressed by

$$\frac{X_1 - M_1}{\sigma_1} = \frac{X_2 - M_2}{\sigma_2} = \frac{X_3 - M_3}{\sigma_3} = \dots$$

These expressions designate deviation measures expressed in terms of the standard deviation as the unit.² Such measures are designated as "standard" and represented by the symbol z . Since the standard deviation is a relatively large unit, decimals will be required in expressing precise standard measures and in the case of those less than the mean, negative numbers will be necessary. In order to avoid these rather inconvenient features, a transformation is usually made to an arbitrary scale which has a convenient mean and whose unit is a fraction of the standard deviation. A convenient scale is one whose mean is 50 and whose standard deviation is 14.³ Inserting these values and using X' to designate the transmuted scores, we have

$$\frac{X'_1 - 50}{14} = \frac{X_1 - M_1}{\sigma_1}$$

$$X'_1 = 50 + \frac{14}{\sigma_1}(X_1 - M_1)$$

Similar formulae may be obtained for other sets of measures merely by changing the subscripts.

The T-scores proposed by McCall⁴ are based upon this principle. He chose a scale of 100 units with the mean at 50 and with a standard deviation of 10 for converting the scores of an unselected group of twelve-year-old pupils. Worlton⁵ has de-

¹ For an illustration in which the median and quartile deviation are used, see Heilman, J. D. "The Translation of Scores into Grades," *Journal of Educational Psychology*, 24: 241-56, April, 1933.

² This procedure was proposed by Woodworth in 1912.

Woodworth, R. S. "Combining the Results of Several Tests, A Study in Statistical Method," *Psychological Review*, 19: 97-123, March, 1912.

³ See Hull, C. L. "The Conversion of Test Scores into Series Which Have Any Assigned Mean and Degree of Dispersion," *Journal of Applied Psychology*, 6: 298-300, September, 1922.

⁴ McCall, W. A. *How to Measure in Education*. New York: The Macmillan Company, 1922, pp. 272-306.

⁵ Worlton, J. T. "The Sigma Index Score as a Standard Measuring Unit," *Elementary School Journal*, 30: 354-62, January, 1930.

scribed the sigma index score which is based on a scale with the mean at 100 and with a standard deviation of 20.

Some workers have transformed scores into *percentile ranks*.¹ The percentile rank of a measure is the per cent of cases in the distribution below the given measure. Its approximate value is that of the nearest percentile point and may be obtained by calculating this point. Precise values are given by the formula:

$$R_x = \frac{100[f(X - l) + S_l]}{Ni}$$

Where R_x is the percentile rank of score X , f is the frequency of the interval in which X occurs, l is the lower limit of this interval, S_l is the number of cases below this lower limit, i is the number of scale units in the interval, and N is the total number of cases.² Usually, the approximate determination mentioned above is sufficiently precise.³ If the fifth, tenth, fifteenth, etc., percentile points are calculated, the percentile ranks for intermediate scores may be found by interpolation. Percentile values may also be obtained graphically from the cumulative frequency curve or from the ogive or percentile curve. These techniques, however, do not give highly precise results.

The unit divisions of the percentile scale are not equal. Those at the extremes are larger than those near the center of the distribution. Also, irregularities in the shape of the distribution tend to produce variations in the scale. Since distributions of raw scores usually present some abnormalities, due to imperfections in the measuring instrument or to selection of the group tested, it has been proposed that the distributions of raw scores be transformed into the normal shape.⁴ The method, however,

¹ For a distinction between "standard percentile ranks" and "centile ranks" see Rogers, D. C. "An Argument for Centile Ranks," *Journal of Educational Psychology*, 24: 107-17, February, 1933.

² Holzinger, *op. cit.*, p. 138.

³ For a method of machine calculation for percentile ranks, see Thurstone, L. L. "Note on the Calculation of Percentile Ranks," *Journal of Educational Psychology*, 18: 617-20, December, 1927.

⁴ Horst, Paul. "A Method for Transforming Any Unimodal Frequency Distribution into a Normal Distribution," *Journal of Educational Psychology*, 24: 129-39, February, 1933.
(Continued next page)

should not be applied when the number of cases is less than three or four hundred and when its use is appropriate, the labor involved will not often be justified. See Chapter VIII for further discussion of percentile ranks and points.

Measures expressed as age scores are considered comparable. Intelligence quotients are also generally considered comparable, but Miller ¹ and Kefauver ² have shown that they are not. The latter gives a table from which equivalent values for certain tests may be read.

Sometimes comparable measures may be secured by calculating ratios. For example, if an investigator desires to compare pupils with reference to the errors made in a set of compositions that vary in length, the number of errors per 100 words may be calculated. If it is desired to compare a number of schools with reference to increase in enrollment, ratios commonly expressed as per cents are usually calculated as a means of making a comparison. This method is sound if the several bases are expressed from absolute zero points. This would be true in the case of the two illustrations just noted. If gains in test scores are being compared, ratios may be used, provided they are labeled "per cent of increase in test scores." If the ratios are labeled or interpreted as "per cent of increase in achievement," they probably are not comparable because the measures of achievement are not expressed from absolute zero points.

D. STATISTICS OF RELATIONSHIP

Techniques for studying the relationship between two sets of paired measures. Two sets of measures, such as the chronological ages and intelligence quotients of a group of students or average marks in high school and average marks in college for

Horst, Paul. "Comparable Scores from Skewed Distributions," *Journal of Experimental Psychology*, 15: 465-68, August, 1932.

Horst, Paul. "A Routine Procedure for Obtaining Comparable Scores," *Journal of Applied Psychology*, 16: 324-30, June, 1932.

¹ Miller, W. S. "The Variation and Significance of Intelligence Quotients Obtained from Group Tests," *Journal of Educational Psychology*, 15: 359-66, September, 1924.

² Kefauver, G. N. "Need of Equating Intelligence Quotients Obtained from Group Tests," *Journal of Educational Research*, 19: 92-101, February, 1929.

the same students are said to be paired because there are two measures for each individual. The degree of relationship refers to the degree to which the larger measures in one set are paired with the larger in the other and the smaller in the one are paired with the smaller in the other, or the degree to which the smaller measures in one set are paired with the larger measures and the larger in the one are paired with the smaller in the other. The relationship in the first type of pairing is designated as positive and that in the second as negative or inverse.

The degree of relationship between two sets of paired data may be studied in several ways. One of the simplest is to write the measures in parallel columns placing the paired items opposite each other. Such an arrangement will be more meaningful if the measures in one column are arranged in order of magnitude. This procedure is crude and when the number of pairs is large, it is more helpful to make a tabulation, as shown in Table IV. Such a correlation table may be partially summarized by calculating the central tendency for each of the columns or each of the rows. The central tendencies thus obtained may be represented graphically as ordinates using the scale divisions of the other dimension of the table as abscissas. The line that "best fits" the several points thus located, summarizes the relationship between the two sets of paired data.¹ One may also prepare a scatter diagram by locating the points corresponding to the several pairs of measures. For some purposes a graphical representation of the averages of the several columns or rows, or a scatter diagram affords a very helpful means of studying relationship, but usually a coefficient of correlation should be calculated as an index of the degree of relationship.

The calculation of the product-moment coefficient of correlation.² The product-moment coefficient of correlation developed by Pearson is the most widely used numerical index of relation-

¹ An illustration of this type of representation is given in Chapter X.

² The critical reader will be interested in the exposition of the assumptions underlying this technique. See pages 101-03.

TABLE IV. ILLUSTRATING A CORRELATION TABLE

First set of measures																
Second set of measures	0	5	10	15	20	25	30	35	40	45	50	55	60	65	T	
												1	1	1	3	
										1		1			2	
								1			1	1			3	
								2	3	2	3				5	
						2	1	4	3	2	1	1			10	
						3	3	2	3	3					13	
			1	1	2	3	3	2	1	1	1				14	
							3	3	1						4	
					1	3	2	2							8	
					3		1								4	
			3	2				1							6	
			2												3	
		1													1	
	T		1	6	7	5	9	11	10	6	9	6	4	1	1	76
M		7.5	19.2	23.9	31.5	31.9	33.9	43.0	45.0	47.5	50.0	58.8	67.5	67.5		

ship. It may be thought of as being defined ¹ by the formula

$$r_{12} = \frac{\sum \frac{x_1 x_2}{\sigma_1 \sigma_2}}{N}$$

in which x_1 and x_2 represent the two sets of paired measures, each measure being expressed as a deviation from the mean of its distribution, σ_1 and σ_2 represent the standard deviations of the two distributions, and N designates the number of pairs of data, or the number of cases. The fraction $\frac{x_1}{\sigma_1}$ is the ratio of a measure to the standard deviation of its distribution. In other

¹ This definition is offered as a convenient approach to the study of correlation. Another approach is to define the coefficient of correlation as the constant r in the linear regression equation. See page 100. It may also be defined as one of the essential indices of the bivariate normal surface. For a brief historical account of the development of correlation and references to original sources, see Walker, H. M. *Studies in the History of Statistical Method*. Baltimore: The Williams and Wilkins Company, 1929, Chapter V. The reader interested in making intensive study of the product-moment coefficient of correlation will find the following reference and the bibliography it includes helpful. Furfey, P. H., and Daly, J. F. "The Interpretation of the Product-Moment Correlation Coefficient," *Catholic University of America Educational Research Monographs*, Vol. 8, No. 4. Washington, D. C.: Catholic University Press, 1934. 57 pp.

words, this fraction represents the measure expressed in terms of the standard deviation as a unit. Hence, the numerator of the formula represents the sum of the products of the paired measures when each is expressed in terms of the standard deviation of its distribution as a unit. The formula may be written

$$r_{12} = \frac{\sum z_1 z_2}{N}$$

If the measures are expressed as deviations from assumed means, as is usually the case when the calculations are made from a correlation table, the following formula ¹ indicates the plan of calculation,

$$\begin{aligned} r_{12} &= \frac{\frac{\sum x'_1 x'_2}{N} - c_1 c_2}{\sqrt{\frac{\sum f_1 (x'_1)^2}{N} - c_1^2} \sqrt{\frac{\sum f_2 (x'_2)^2}{N} - c_2^2}} \\ &= \frac{\frac{\sum x'_1 x'_2}{N} - c_1 c_2}{\sigma_1 \sigma_2} \end{aligned}$$

¹ Several other forms of the formula may be written. See Dunlap, J. W., and Kurtz, A. K. *Handbook of Statistical Nomographs, Tables, and Formulas*. Yonkers-on-Hudson: World Book Company, 1932, pp. 117-19. Fifty-two forms are given by Symonds, P. M. "Variations of the Product Moment (Pearson) Coefficient of Correlation," *Journal of Educational Psychology*, 17: 458-69, October, 1926.

A formula developed by Pearson for calculating the product-moment coefficient utilizing the means of the columns or of the rows is given by Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, pp. 258-60. This formula may be expressed as

$$r_{12} = \frac{\sum f_i x'_1 (M_i - M_2)}{N \sigma_1 \sigma_2}$$

or

$$r_{12} = \frac{\sum f_i x'_2 (M_i - M_1)}{N \sigma_1 \sigma_2}$$

In the first formula M_i refers to the means of the columns and M_2 to the mean of the total distribution of X_2 , assuming that X_2 is distributed along the vertical axis of the correlation table as in Figure 1. The numerator is then the sum of the products of the differences and the corresponding totals of the columns and deviations of the columns from the assumed mean of X_1 . Corresponding statements apply to the second formula. When employing either of these formulae, all quantities should be expressed in terms of scale units rather than intervals.

The application of this formula to a correlation table is illustrated in Figure 1. Heavy horizontal and vertical lines have been drawn to mark off the scale interval in each distribution within which the mean is assumed to lie. The correlation table is extended at the right and at the bottom to provide for the following quantities: deviations from the assumed means designated by the symbols x'_2 and x'_1 , products of each deviation by its frequency expressed by the symbols $f_2x'_2$ and $f_1x'_1$, the products of each frequency and the square of its deviation expressed by the symbols $f_2(x'_2)^2$ and $f_1(x'_1)^2$. From the algebraic sums of the appropriate ones of these columns and rows, the standard deviations σ_1 and σ_2 are calculated¹ according to the method illustrated on page 76. Two additional columns are given at the right of the table. The values under the symbol $\Sigma x'_1$ are obtained by multiplying the measures in each horizontal row or array by their corresponding x'_1 deviations. For example, $1 \times 4 = 4$ and $(1 \times 2) + (2 \times 5) + (1 \times 7) = 19$. The lowest horizontal row (to the right of 30) yields $(1 \times -5) + (1 \times -3) = -8$. The values under the symbol $\Sigma x'_1x'_2$ are obtained by multiplying $\Sigma x'_1$ values by their corresponding x'_2 deviations. For example, $4 \times 7 = 28$ and $19 \times 6 = 114$. The horizontal rows following the symbols $\Sigma x'_2$ and $\Sigma x'_1x'_2$ provide checks for the sums of the columns labeled $f_2x'_2$ and $\Sigma x'_1x'_2$. The values in these rows are obtained similarly to those in columns $\Sigma x'_1$ and $\Sigma x'_1x'_2$. For example, $(1 \times -6) + (1 \times -8) = -14$ which multiplied by -5 equals $+70$. The sum of the column $\Sigma x'_1$ should also be the same as the sum of the horizontal row $f_1x'_1$. Thus, in the illustration given, the values 43, -348 , and 2333, later used in calculation, are each obtained twice.² Furthermore, the value for N , or 568, may be checked

¹ It should be noted that the standard deviations are expressed in class intervals as units, rather than in scale units. For purposes *other than the calculation of r_{12}* , the standard deviations should be expressed in scale units. In the case of the correlation table of Figure 1 each of the values should be multiplied by 5.0, the width of the class or scale interval.

² Through a mere coincidence, the sum of the negative values in the $\Sigma x'_1$ column equals -348 . The checks mentioned refer to the $f_1x'_1$ row and the $\Sigma x'_1$ column (43), the $\Sigma x'_2$ row and the $f_2x'_2$ column (-348), and the $\Sigma x'_1x'_2$ row and column (2333). No check is given for the $f_1(x'_1)^2$ row and the $f_2(x'_2)^2$ column.

OTIS SCORE, X_1																				
	0	5	10	15	20	25	30	35	40	45	50	55	60	f_2	x'_2	$f_2x'_2$	$f_2(x'_2)^2$	$\Sigma x'_1$	$\Sigma x'_1x'_2$	
105															1	7	7	49	4	28
100								1			1		1		4	6	24	144	19	114
95							1	1	1	5	2	1			11	5	55	275	42	210
90							2	4	7	4	3				20	4	80	320	62	248
85		1					7	6	12	6	2				34	3	102	306	85	255
80					1	1	6	8	16	7	1	1			42	2	84	168	79	158
75					3	3	16	24	10	4					57	1	57	57	53	53
70			1	7	13	32	30	12	5	2					102	0	+409	0	47	0
65		1	1	10	26	26	23	6	1						94	-1	-94	94	-15	15
60		3	6	19	29	15	3	2							77	-2	-154	308	-90	180
55		1	10	14	15	8	3	1							52	-3	-156	468	-72	216
50		1	12	8	14	2	1								38	-4	-152	608	-69	276
45		5	8	4	4										21	-5	-105	525	-56	280
40	1	1	3	6											11	-6	-66	396	-30	180
35		2													2	-7	-14	98	-8	56
30	1		1												2	-8	-16	128	-8	64
f_1	2	15	42	68	105	105	102	59	37	20	10	2	1	568		-757	3944	+391	2333	
x'_1	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7			+409		-348		
$f_1x'_1$	-10	-60	-126	-136	-105	-437	102	118	111	80	50	12	7			-348		43		
$f_1(x'_1)^2$	50	240	378	272	105	0	102	236	333	320	250	72	49				480 - 437 = 43			
$\Sigma x'_2$	-14	-56	-157	-178	-200	-60	32	74	86	70	42	7	6			2407		- 665 + 317 = - 348		
$\Sigma x'_1x'_2$	70	224	471	356	200	0	32	148	238	280	210	42	42			2333				

$$c_1 = \frac{43}{568} = .0757 \quad c_1^2 = .0057$$

$$c_1 c_2 = -.0464$$

$$c_2 = \frac{-348}{568} = -.6127 \quad c_2^2 = .3754$$

$$\sigma_1 = \sqrt{\frac{\Sigma f_1(x'_1)^2}{N} - c_1^2} = \sqrt{\frac{2407}{568} - .0057} = 2.0572$$

$$\sigma_2 = \sqrt{\frac{\Sigma f_2(x'_2)^2}{N} - c_2^2} = \sqrt{\frac{3944}{568} - .3754} = 2.5629$$

$$r_{12} = \frac{\frac{\Sigma x'_1 x'_2}{N} - c_1 c_2}{\sigma_1 \sigma_2} = \frac{\frac{2333}{568} - (-.0464)}{2.0572 \times 2.5629} = .7878 = .79$$

FIG. 1. These formulae and the facing chart show the calculation of a Pearson product-moment coefficient of correlation.

by totaling the f_2 column and the f_1 row. The student should take advantage of these checks and should not commence the later steps in the calculation of the coefficient of correlation until he has made them.

The calculation of c_1 , c_2 , σ_1 , and σ_2 requires no explanation. The sign of c_1^2 and c_2^2 is always positive. The sign of c_1c_2 , however, should be carefully noted. In the illustration it is negative, since c_1 and c_2 are unlike in sign. The reader should observe that calculations were carried to the fourth decimal place. This facilitates rounding off correctly to two places. Reporting a coefficient to more than two decimal places lends a false air of accuracy which should be avoided.¹

Economy in calculating a coefficient of correlation. The construction of a correlation table and the calculation of the coefficient from it is a laborious process, especially when the number of cases is large. Several statistical workers have devised blank forms to facilitate and routinize the process.² The use of a wisely planned correlation chart not only results in considerable saving of time, but it also adds to the accuracy of the computations by systematizing the work. Greater economies, however, may be effected in other ways.

The tabulation of the data in a correlation table is not necessary and the elimination of this step may result in considerable saving of time, especially when the intercorrelations between

¹ See page 62.

² Holzinger, K. J. *Statistical Methods for Students of Education*. Boston: Ginn and Company, 1928, p. 155.

Justice, W. A. "Correlation Sheet." Cincinnati: C. A. Gregory Company.

Kelley, T. L. "Kelley Correlation Chart." Yonkers-on-Hudson, New York: World Book Company. (A copy is pasted on the inside of the back cover of Kelley's *Interpretation of Educational Measurements*.)

Lauer, A. R. "Simplex Correlation Form." Minneapolis: The Educational Test Bureau.

Otis, A. S. "The Otis Correlation Chart," *Journal of Educational Research*, 8: 440-48, December, 1923.

Otis, A. S. *Statistical Method in Educational Measurement*. Yonkers-on-Hudson, New York: World Book Company, 1925, p. 195.

Thurstone, L. L. "A Data Sheet for the Pearson Correlation Coefficient," *Journal of Educational Research*, 6: 49-56, June, 1922.

Toops, H. A. "A Printed Form for Computing the Standard Deviation on the Adding Machine," *Journal of Educational Research*, 12: 56-58, June, 1925.

several paired variables are desired. Walker ¹ has proposed a procedure, the first step of which is to assemble the values of each variable in a frequency distribution and then to calculate the standard deviations. The paired measures are written in parallel columns with space between the columns for writing in the corresponding deviations in terms of intervals from the assumed means. The products of the paired deviations are then written in two columns, one for the positive values and one for the negative. The algebraic sum of the totals of these columns will be the term $\Sigma x'_1 x'_2$. The corrections obtained in calculating the standard deviations are those required for completing the formula

$$r_{12} = \frac{\frac{\Sigma x'_1 x'_2}{N} - c_1 c_2}{\sigma_1 \sigma_2}$$

When a suitable calculating machine is available, another form of the formula affords a convenient procedure. If $\sqrt{\frac{\Sigma x_1^2}{N}}$ and $\sqrt{\frac{\Sigma x_2^2}{N}}$ are substituted for σ_1 and σ_2 , we obtain

$$r_{12} = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2} \sqrt{\Sigma x_2^2}}$$

Also $x_1 = X_1 - M_1$ and $x_2 = X_2 - M_2$. Substituting these equivalents

$$\begin{aligned} r_{12} &= \frac{\Sigma (X_1 - M_1)(X_2 - M_2)}{\sqrt{\Sigma (X_1 - M_1)^2} \sqrt{\Sigma (X_2 - M_2)^2}} \\ &= \frac{\Sigma X_1 X_2 - \Sigma X_1 M_2 - \Sigma X_2 M_1 + \Sigma M_1 M_2}{\sqrt{\Sigma X_1^2 - 2\Sigma X_1 M_1 + \Sigma M_1^2} \sqrt{\Sigma X_2^2 - 2\Sigma X_2 M_2 + \Sigma M_2^2}} \end{aligned}$$

M_2 being a constant, $\Sigma X_1 M_2 = M_2 \Sigma X_1$. Since $M_1 = \frac{\Sigma X_1}{N}$, $\Sigma X_1 = N M_1$. Hence, $M_2 \Sigma X_1 = M_2 (N M_1)$. $\Sigma X_2 M_1 = M_1 \Sigma X_2$

¹ Walker, J. F. "Short Method for Finding Zero Order Coefficients of Correlation," *Journal of Educational Psychology*, 21: 65-67, January, 1930.

$= M_1(NM_2)$. Since M_1 and M_2 are constants, $\Sigma M_1 M_2$ is equal to $NM_1 M_2$, ΣM_1^2 is equal to NM_1^2 and ΣM_2^2 is equal to NM_2^2 . Similarly, $\Sigma X_1 M_1 = NM_1^2$ and $\Sigma X_2 M_2 = NM_2^2$. Substituting these equivalents, the equation simplifies to the form

$$r_{12} = \frac{\Sigma X_1 X_2 - NM_1 M_2}{\sqrt{\Sigma X_1^2 - NM_1^2} \sqrt{\Sigma X_2^2 - NM_2^2}}$$

The calculations indicated by this formula may be made from a correlation table,¹ but unless N is very large, it is more economical to make them directly from the data records,² that is, from the series of paired measures of X_1 and X_2 . However, if the measures are expressed in terms of large numbers, the calculations will be cumbersome, even when machines are available for making them. The labor involved can be reduced by transmuting the measures into smaller numbers.³ The smallest measure of each series may be subtracted from all the other measures in that series, i.e., the smallest value of X_1 may be subtracted from all the values of X_1 , and the smallest value of X_2 may be subtracted from all the values of X_2 . It is not essential in using this formula that the measures in either series be arranged in order of magnitude. It is only essential that the measures in the original or in the reduced series remain correctly paired. Another plan is to form a scale of intervals

¹ If standard deviations are desired in addition to r_{12} , divide numerical values of $\sqrt{\Sigma X_1^2 - NM_1^2}$ and $\sqrt{\Sigma X_2^2 - NM_2^2}$ by N before extracting square roots.

For a description of the technique, see Ackerson, Luton. "A Pearson-r Form for Use with Calculating Machines," *Journal of Educational Psychology*, 19: 58-60, January, 1928.

² For an illustration see Ayres, L. P. "Shorter Method for Computing the Coefficient of Correlation," *Journal of Educational Research*, 1: 216-21, March, 1920.

For detailed instructions for using a Monroe Calculating Machine with a minor modification of this formula, see Tremmel, E. E., and Weidemann, C. C. "A Machine Method of Calculating the Pearson Correlation Coefficient," *University of Nebraska Publication*, No. 72. Lincoln, Nebraska: Extension Division, University of Nebraska, June, 1930. 15 pp.

³ Ayres, L. P. "Substituting Small Numbers for Large Ones in the Computation of Coefficients of Correlation," *Journal of Educational Research*, 2: 502-04, June, 1920.

Toops, Herbert A. "Computing Intercorrelations of Tests on the Adding Machine," *Journal of Applied Psychology*, 6: 172-84, June, 1922.

such as would be used if the data were to be tabulated in a correlation table. The first interval may be given a value of 1, the second a value of 2, and so on. These values may be substituted for the measures in the respective intervals. The process is called coding.

When making calculations from untabulated measures, one should be systematic. It is helpful to list the measures, their squares, and their products under the following heads: ΣX_1 , ΣX_2 , ΣX_1^2 , ΣX_2^2 , and $\Sigma X_1 X_2$, with the totals appearing at the bottoms of the columns. The means M_1 and M_2 are obtained by dividing ΣX_1 and ΣX_2 by N . Where several coefficients are to be calculated from a number of series referring to the same individuals, the table may be extended to include the following heads: ΣX_1 , ΣX_2 , $\Sigma X_3 \dots \Sigma X_n$; ΣX_1^2 , ΣX_2^2 , ΣX_3^2 , \dots ΣX_n^2 ; $\Sigma X_1 X_2$, $\Sigma X_1 X_3$, $\Sigma X_2 X_3 \dots \Sigma X_{n-1} X_n$. This procedure eliminates duplication of calculation.

The term $\Sigma X_1 X_2$ is inconvenient to compute, even with the aid of a calculating machine, but it may be eliminated by a further transformation of the formula for the coefficient of correlation.¹

$$r_{12} = \frac{\sigma_1}{2\sigma_2} + \frac{\sigma_2}{2\sigma_1} - \frac{\frac{\Sigma D^2}{N} - (M_1 - M_2)^2}{2\sigma_1\sigma_2}$$

This formula, in which D designates the difference between the measures of a pair appears formidable, but examination of it will reveal that the calculations called for are simple. If the standard deviations and the means are desired for other purposes, the calculation of the coefficient of correlation by means

¹ Huffaker, C. L. "A Note on Statistical Methods," *Journal of Educational Psychology*, 16: 265-66, April, 1925.

Orleans, J. S. "Correlation without Plotting," *Journal of Educational Psychology*, 18: 310-17, May, 1927.

Cureton, E. E. "Computation of Correlation Coefficients," *Journal of Educational Psychology*, 20: 588-601, November, 1929.

For a plotting and checking technique based upon a modification of this formula, see

Anderson, L. D., and Toops, H. A. "A New Apparatus for Plotting and a Checking Method for Solving Large Numbers of Intercorrelations," *Journal of Educational Psychology*, 19: 650-57, December, 1928; 20: 36-43, January, 1929.

of this formula requires only the summation of the squares of the differences and the other indicated operations. If one set of the raw measures is transmuted so that $\sigma_1 = \sigma_2$ and $M_1 = M_2$, the formula becomes

$$r_{12} = 1 - \frac{\frac{\sum D^2}{N}}{2\sigma^2}$$

Feldstein¹ has described the technique for carrying out the operations required by this formula. He reports that a relatively inexperienced statistical worker computed the forty-five intercorrelations for ten variables, $N = 100$, in a total of six hundred thirty-five minutes. This is an average of about fourteen minutes per coefficient of correlation.

In some situations the calculation of the tetrachoric coefficient of correlation may advantageously be substituted for the calculation of the product-moment coefficient of correlation. For the calculation of a tetrachoric coefficient, the data are tabulated in a 2×2 table and under certain conditions it is equivalent to the product-moment coefficient of correlation.²

The computation of coefficients of correlation from tabulating machine cards has been described by Mendenhall and Warren.³ A machine of somewhat different type has been constructed by Hull.⁴

Although the techniques referred to in the preceding paragraphs are economical of time, the elimination of the correlation table from the process may be a handicap in interpreting

¹ Feldstein, M. J. "A New Technique for Machine Computation of Coefficients of Correlation," *Journal of Experimental Education*, 2: 278-82, March, 1934.

² The calculation of tetrachoric coefficients of correlation is most readily accomplished through the use of diagrams contained in the following publication:

Cheshire, Leone; Saffir, Milton; and Thurstone, L. L. *Computing Diagrams for the Tetrachoric Correlation Coefficient*. Chicago: University of Chicago Bookstore, 1933.

³ Mendenhall, R. M., and Warren, Richard. "Computing Statistical Coefficients from Punched Cards," *Journal of Educational Psychology*, 21: 53-62, January, 1930.

⁴ Hull, C. L. "An Automatic Correlation Calculating Machine," *Journal of the American Statistical Association*, 20: 522-31, December, 1925.

the coefficient of correlation as an index of the relationship between the two sets of paired measures. Furthermore, the correlation table is useful as a means of spotting probable cases of non-linear relationship to which the product-moment coefficient is not applicable. Hence, the elimination of the correlation table is not to be recommended, except when the saving of time is an important consideration. In general, the investigator who computes coefficients of correlation only occasionally, should follow the procedure illustrated on pages 90-91 or some minor modification of it.

Error due to grouping in broad categories. In the calculation from the correlation table on page 90, the mid-point of an interval was taken as the average value of the measures in it. In a normal distribution the mean value of the measures in an interval is slightly nearer the mean of the distribution than the mid-point of the interval. Hence, the standard deviations in the denominator of the formula for the coefficient of correlation will be slightly too large and the value of r will be decreased. Sheppard's correction¹ may be applied but if the correlation table is not less than 12×12 or 10×10 , the error is relatively small.² For a table smaller than 10×10 , the error due to broad categories or intervals is large enough to be a matter of considerable concern, especially when N is less than 50. When the coefficient of correlation is calculated between two sets of school marks or other measures expressed in terms of a small number of categories by the method described on pages 90-91, an investigator should recognize that the result is likely to involve a relatively large error. In such cases the method of tetrachoric correlation or polychoric correlation may be used.³

¹ See page 154.

² For a 10×10 table, the error is about 4 per cent.

Camp, B. H. *The Mathematical Part of Elementary Statistics*. Boston: D. C. Heath and Company, 1931, p. 301.

See also Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, pp. 167 f.

³ Camp, *op. cit.*, pp. 302 f.

See also Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, p. 263.

Studying the relationship between two sets of paired measures when it is non-linear. The derivation of the formula for the product-moment coefficient of correlation is based on the assumption that the relationship is linear, which means that the line joining the points located in a graphical representation of the means of the columns or the rows approaches a straight line, rather than a section of a parabola or some other curve. When the relationship is non-linear, the correlation ratio should be calculated. When the calculation is made from a correlation table, the following formulae indicate the procedure.¹

$$\eta_{12} = \frac{\sqrt{\frac{\sum (\Sigma x'_1)^2}{f_2} - c_1^2}}{\sigma_1} \quad \begin{array}{l} \text{(Correlation ratio for means of} \\ \text{arrays or rows, curvilinear cor-} \\ \text{relation of } x_1 \text{ on } x_2) \end{array}$$

$$\eta_{21} = \frac{\sqrt{\frac{\sum (\Sigma x'_2)^2}{f_1} - c_2^2}}{\sigma_2} \quad \begin{array}{l} \text{(Correlation ratio for means of} \\ \text{columns, curvilinear correlation of} \\ \text{ } x_2 \text{ on } x_1) \end{array}$$

In calculating the ratios of correlation by means of these formulae, two columns and two rows are added to the correlation table. (See page 90.) The first of the columns is headed $(\Sigma x'_1)^2$ and its values are obtained by squaring those appearing in the $\Sigma x'_1$ column. The second of the two new columns is headed $\frac{(\Sigma x'_1)^2}{f_2}$. Its values are obtained by dividing the values in the $(\Sigma x'_1)^2$ column by the corresponding values in the f_2 column. The new horizontal rows, appearing at the bottom of the table, $(\Sigma x'_2)^2$ and $\frac{(\Sigma x'_2)^2}{f_1}$, are obtained similarly. The final steps in the calculation of the ratios should be evident from the formulae.²

¹ See page 252 for a different form of these formulae.

² The reader who desires additional information will find it helpful to consult:

Odell, C. W. *Educational Statistics*. New York: The Century Company, 1925, pp. 209 f. (Continued next page)

The coefficient of correlation may be thought of as measuring the deviations from the straight line which best fits the means of the columns or rows. The correlation ratio may be thought of as measuring the deviations from the curved line which best fits the means of the rows or columns. It should be noted that for any tabulation there are two correlation ratios—one that relates to the curve best fitting the means of the columns and the other relating to the curve that best fits the means of the rows.

The regression equation as an expression of relationship. On page 86 the suggestion was made that the relationship between two sets of paired measures might be summarized by representing means of the columns of the correlation table graphically as ordinates above the mid-points of the intervals of the horizontal scale of the table and then drawing the straight line that "best fits" the points thus located. For some purposes it is desirable to determine the equation of this line which will express the relation between the "average" value of the measures of one set associated with a given measure of the other set. If \bar{X}_1 is employed to represent these "average" values, the form of the equation will be

$$\bar{X}_1 = mX_2 + C$$

In general the corresponding values of X_1 and \bar{X}_1 will not be equal and the "best fitting" line is defined as the one for which $\Sigma(X_1 - \bar{X}_1)^2$ is a minimum. When the relationship is linear, the equation of this line, called the *regression equation*,¹ is

Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, Chapter X.

A simplified method of calculation has been reported by Dvorak. Dvorak, August. "A Simplified Computation of Non-Linear Correlation," *Journal of Educational Research*, 25: 99-104, February, 1932. A chart to facilitate this simplified method of calculation is published by Longmans, Green and Company, New York.

¹ For an explanation of the derivation of this equation see Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, pp. 158 f.

Use of the word regression to designate this equation is due to Francis Galton who studied the inheritance of traits and found that offspring tend to resemble the mid-parent, "average of the measures of father and mother." This tendency was called regression. When the relationship was expressed in equation form,

$$\bar{X}_1 = r_{12} \frac{\sigma_1}{\sigma_2} X_2 - r_{12} \frac{\sigma_1}{\sigma_2} M_2 + M_1$$

This equation may be written in the following forms

$$\frac{\bar{X}_1 - M_1}{\sigma_1} = r_{12} \frac{X_2 - M_2}{\sigma_2}$$

$$\bar{X}_1 - M_1 = r_{12} \frac{\sigma_1}{\sigma_2} (X_2 - M_2)$$

The relation between the "average" values of X_2 and the corresponding values of X_1 is given by the equation

$$\bar{X}_2 = r_{12} \frac{\sigma_2}{\sigma_1} X_1 - r_{12} \frac{\sigma_2}{\sigma_1} M_1 + M_2$$

The use of the regression equation as a formula for prediction is dealt with in Chapter X. Multiple correlation and multiple regression are also considered therein.

Techniques for the study of relationships involving other types of data.¹ In the preceding descriptions, both sets of data have been expressed in quantitative terms. One or both sets of the data may be rank orders or categorical classifications. If a normal distribution is assumed, ranks may be transformed into quantitative measures and the product-moment coefficient of correlation or the correlation ratio is then an appropriate

the term regression was used to identify it. For a brief account of Galton's work see Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, pp. 152 f.

¹ The limitations of space prohibit the description of these techniques. The interested reader may consult the following references.

Kelley, T. L. *Statistical Method in Education*. New York: The Macmillan Company, 1923, pp. 231-78.

Holinger, K. J. *Statistical Methods in Education*. Boston: Ginn and Company, 1928, pp. 231-78.

Walker, Helen M. *Studies in the History of Statistical Method*. Baltimore: The Williams and Wilkins Company, 1929, pp. 125-41. This reference consists mainly of an annotated bibliography.

These references describe certain techniques in addition to those mentioned here. Dunlap, J. W., and Kurtz, A. K. *Handbook of Statistical Nomographs, Tables, and Formulas*. Yonkers-on-Hudson, New York: World Book Company, 1932, pp. 122-25 may be consulted for formulae.

technique. When both sets of data are in terms of ranks, a measure of the relationship may be calculated by the formula

$$\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

in which d represents the difference in the rank of the two measures of a pair. Measures in quantitative terms may be transformed into ranks and this formula applied. Values of ρ (rho) differ slightly from the corresponding values of r .¹

If one set of data is in terms of quantitative measures and the other is in terms of categorical classifications such as grade levels, geographical areas, types of school organization, and the like, the correlation ratio may be calculated by means of a formula given in Chapter VIII, page 252. If there are only two categories (dichotomous classification), *bi-serial* r is used as the measure of relationship. See page 236 for formula.

If one set of data consists of ranks and the other of categorical classifications, the ranks may be transformed into quantitative measures if the assumption of a normal distribution appears justified. When this is done, the correlation ratio may be calculated. If both sets of data are categorical classifications, the *coefficient of contingency* gives a measure of the relationship. If both of the classifications are dichotomous, the *tetrachoric coefficient of correlation* may be calculated.

Assumptions relating to the calculation of the coefficient of correlation.² The assumptions relating to the product-moment coefficient of correlation can best be understood if one thinks of the usual correlation table. The only assumption made in the derivation of the formula is that the distribution of the frequencies in this table exhibit a linear relationship rather than one that

¹ The relationship is $r = 2 \sin \frac{\pi}{6} \rho$.

Kelley, *op. cit.*, p. 193, gives a table of the corresponding values.

² For a report of a study of the extent to which assumptions are ignored in practice, see

Furfey, P. H., and Daly, J. F. "Product-Moment Correlation as a Research Technique in Education," *Journal of Educational Psychology*, 26: 206-11, March, 1935.

is curvilinear. Cases of marked non-linearity may be identified from an inspection of the correlation table, but for precise determinations the Blakeman test¹ should be applied and if a satisfactory degree of linearity is not shown, the product-moment coefficient should not be used.

When a coefficient of correlation is calculated from a correlation table, the assumption is made that the measures falling within the several cells of the table are uniformly distributed within the areas. This assumption is not completely satisfied when the distributions are normal, and frequently the departure from it is much greater, especially when N is not large. The effect of non-conformity with this assumption is indicated by the fact that different choices of intervals for the correlation table may yield coefficients varying as much as .10 or more.² Sheppard's correction for coarse grouping will usually give a more correct coefficient³ but unless N is greater than 100, a better procedure is to make the calculations directly from the data by means of the formula on page 94.

In our use of the standard error of estimate⁴ two additional assumptions are introduced. The first of these requirements is homoscedasticity, which means that the variabilities (standard deviations) of the several arrays (columns and rows) of the cor-

¹ The Blakeman test for linearity is based on the difference between the correlation ratio (η) and the coefficient of correlation (r). The expression to be evaluated may be written

$$N(\eta^2 - r^2)$$

When the relationship is perfectly linear, the value of this expression is zero, and as the relationship becomes curvilinear, the value of the expression indicates the degree of departure from linearity. Although there is not complete agreement between authorities, the product-moment coefficient probably can be safely used when

$$N(\eta^2 - r^2) < 16.38$$

A more conservative limit is 11.37 and it is wise to use this limit when N is not large. For a longer form of Blakeman's test see Holzinger, *op. cit.*, p. 183.

² Lauer, A. R. "An Empirical Study of the Effects of Grouping Data and Calculation of r by the Pearson Product-Moment Method," *Journal of Applied Psychology*, 14: 182-89, April, 1930. This reference gives the results of computing coefficients of correlation from different groupings of the same data. It includes an excellent summary statement of the factors which may affect the coefficient of correlation calculated from a given group of data.

³ See pages 153 f.

⁴ See page 333.

relation table are equal. The second is that the correlation surface¹ is normal. Although neither of these requirements is involved in the calculation of r , their introduction in our use of the standard error of estimate has associated them with the coefficient of correlation, and for practical purposes it is wise to regard them as implied assumptions which should be approximated. A coefficient of correlation is thought of as an index descriptive of the correlation table. If the correlation surface is not approximately homoscedastic and normal, the coefficient is likely to be misleading. In an abnormal correlation table a few pairs of measures may affect the coefficient of correlation to such an extent that their omission would materially change the calculated value. The situation is analogous to the use of the mean as a measure of the central tendency of a distribution in which a few extreme measures cause the difference between the mean and the median to be relatively large.

E. MEASURES OF THE EFFECT OF CHANCE IN RANDOM SAMPLING

The probable limits of the value of a statistic for a universe² when calculations have been made from a random sample. An investigator necessarily works with a limited collection or sample of data, but frequently he is interested in the value of certain statistics for a larger population or universe. In other words, he wishes to generalize from his findings. If the sample is a random one,³ it is possible to determine the probable limits of the value of the statistic for the universe from which the sample was taken or is considered to be taken.

¹ The term "correlation surface" implies representation in three dimensions, the frequencies in the correlation table being represented as vertical distances. The correlation surface is normal when all of the arrays, both columns and rows, form normal distributions.

² A universe is defined as an infinitely large population. The formulae given in the following paragraphs are based on the assumption of a random sample from a universe. They, however, may be used with random samples from large finite populations. In such cases the calculated value of the standard or probable error will tend to be slightly too large. This condition is not undesirable because it merely makes the probable limits slightly conservative. See footnote, page 108.

³ See pages 57 f. and 155 f. for discussion of random sampling.

If a series of similar, but independent random samples, are taken from a universe, the values of a statistic, such as the mean, will not be identical, but will tend to form a normal distribution whose mean approaches the value of the statistic for the universe.¹ The standard deviation of the distribution of the means of an infinite number of similar but independent random samples is given by the formula:²

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

The median deviation, commonly designated as the probable error (PE), is usually a more convenient measure of the variability of this distribution. The formula is

$$PE_M = .6745 \frac{\sigma}{\sqrt{N}}$$

For the standard deviation and the coefficient of correlation³ the corresponding formulae are

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} = .7071 \frac{\sigma}{\sqrt{N}}$$

$$PE_\sigma = .6745 \frac{\sigma}{\sqrt{2N}} = .4769 \frac{\sigma}{\sqrt{N}}$$

$$\sigma_r = \frac{1 - r_{12}^2}{\sqrt{N}}$$

$$PE_r = .6745 \frac{1 - r_{12}^2}{\sqrt{N}}$$

¹ For an illustration using a large finite population see Chaddock, R. E. *Principles and Methods of Statistics*. Boston: Houghton Mifflin Company, 1925, pp. 232 f.

² When the data are fallible, i.e., involve variable errors of measurement, this formula gives a measure of the combined effect of sampling and of such errors. See page 157 for formulae for the effect of sampling alone when the data are fallible.

³ When a coefficient of correlation has been corrected for attenuation (see page 151) the correct formula varies with the one used to secure correction for attenuation. See Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, pp. 210 f.

The probable error of a proportion or per cent (p) is ¹

$$PE_p = .6745\sqrt{\frac{pq}{N}}$$

in which q is defined as $1 - p$.

The probable error of the difference between means of two samples is found by the formula ²

$$PE_{D_{1-2}} = \sqrt{\overline{PE}_{M_1}^2 + \overline{PE}_{M_2}^2 - 2r_{12}PE_{M_1}PE_{M_2}}$$

If the two sets of measures are not correlated, i.e., if $r_{12} = 0$, the formula simplifies to

$$PE_{D_{1-2}} = \sqrt{\overline{PE}_{M_1}^2 + \overline{PE}_M^2}$$

The probable error of the difference of two proportions is similar.

The values of a statistic calculated from similar but independent random samples form a normal distribution and hence the relationship between abscissa distances and corresponding areas under the normal curve may be applied. See page 80. Fifty per cent of these values of a statistic would be within $1.00PE$ of the mean of the distribution (the value of the statistic for the universe). Hence, if the value calculated from a single random sample is taken as an estimate of the value of the statistic for the universe, the chances are just even (50 to 50) that the value of the statistic for the universe will be within $\pm 1.00PE$ of the estimate. These limits are commonly expressed by connecting the calculated value with its probable error by a plus or minus sign (\pm). As a means of distinguishing between the

¹ This is the formula usually given. A simple random sampling is assumed. For modified random selection, the formula is different. See Mudgett, B. D., and Gevorkiantz, S. R. "Reliability of Forest Surveys," *Journal of the American Statistical Association*, 29: 257-81, September, 1934.

² This is the formula usually given. In the derivation the coefficient of correlation in the product term appears as $r_{M_1M_2}$, but if the assumptions of random sampling are met, it is equivalent to r_{12} . For proof see Walker, Helen M. "A Note on the Correlation of Averages," *Journal of Educational Psychology*, 19: 636-42, December, 1928.

statistic for the sample and that of the universe, \sim (read curl¹) may be written over the usual symbol. For example, if the mean of a random sample is 57.00 and its probable error is 1.50, we would write $M = 57.00$ and $\hat{M} = 57.00 \pm 1.50$. The expression 57.00 ± 1.50 is not to be interpreted literally but rather as a statement of the probable limits of the mean of the universe. By means of a table of the values of the probability integral, described on page 80, the probabilities for other limits may be determined. The probabilities corresponding to the limits defined by various multiples of the probable error² are given in Table V.

TABLE V. THE PROBABILITIES THAT THE VALUE OF A STATISTIC FOR A UNIVERSE LIES WITHIN THE INTERVAL FORMED BY SUBTRACTING AND ADDING A MULTIPLE OF ITS PROBABLE ERROR

PROBABILITY	INTERVAL
1 to 1	statistic $- PE$ to statistic $+ PE$
4.6 to 1	statistic $-2PE$ to statistic $+2PE$
22 to 1	statistic $-3PE$ to statistic $+3PE$
142 to 1	statistic $-4PE$ to statistic $+4PE$
1340 to 1	statistic $-5PE$ to statistic $+5PE$

The meaning of a difference³ depends upon its sign. Hence, an investigator is interested in ascertaining the probabilities that the difference for the universe has the same sign as the obtained difference. When computations have been made from random samples, the ratio of a difference to its probable error has been proposed as a convenient statistic. For a ratio of 1.00 the chances of the difference for the universe having the same sign are 75 in 100; for a ratio of 1.50, 84 in 100; for a ratio of 2.00, 91 in 100; for a ratio of 3.00, 98 in 100; for a ratio of 4.00, 997 in 1000. The theoretical probabilities for various values of this ratio, commonly called critical ratio (*CR*), have

¹ This symbol is not in general use but the need for it is obvious. It is used in Camp, B. H. *The Mathematical Part of Elementary Statistics*. Boston: D. C. Heath and Company, 1931, p. 241.

² The standard error might be used, but it is less convenient.

³ A similar statement may be made with reference to the coefficient of correlation. See page 117.

been determined and assembled in table form.¹ Such a table affords a means of interpreting a difference obtained from two independent random samples with reference to the difference for the corresponding universes, *provided* the difference is not influenced by systematic errors of measurement or other data faults. When the critical ratio is 4.00 or greater, the difference is commonly called *statistically significant*. This is a technical term and should not be interpreted as meaning that the difference is necessarily dependable or is significant in a practical sense.² For further discussion of the interpretation of *differences*, see pages 237, 244 f., and 305 f.

The probable error of statistics as a measure of reliability. In a number of texts the probable error of statistics is discussed under the head of "reliability."³ The use of this caption tends to be misleading because it represents a half truth. When the data are accurate, a complete statement would be, "reliability of statistics derived from a large random sample of accurate data when used as estimated values of the corresponding universe." When the data are inaccurate, and the type of inaccuracy is that of variable errors of measurement, the probable error formulae account for the effect of chance in the selection of the sample and of the variable errors of measurement. The formulae do not give a measure of the effect of systematic errors of sampling (bias in the selection of the sample), systematic errors of measurement, variable errors of validity, and systematic errors of validity. This point is important. The probable error was developed by astronomers and workers in other sciences who were dealing with groups of data that might be assumed to be random samples of universes. Furthermore, these data did not involve errors of validity, and systematic errors of measurement were disregarded. When the

¹ Garrett, H. E. *Statistics in Psychology and Education*. New York: Longmans, Green and Company, 1926, p. 135.

² For a discussion of this point see Lincoln, E. A. "The Insignificance of Significant Differences," *Journal of Experimental Education*, 2: 288-90, March, 1934.

³ For discussion of reliability see pages 177 and 199 f.

probable error technique was taken over by workers in the field of educational research, they failed to keep in mind the conditions under which the technique was developed.

Another point to be noted is in regard to what the probable error is. Not infrequently one finds in educational writings a statement to the effect that the value of a certain statistic calculated from a sample is a certain magnitude plus or minus its probable error. Such statements are absurd. The mean of a group of data *is* the calculated mean of that group and is to be considered accurate, provided the assumptions underlying the procedure of calculating it are satisfied and no errors were made in the arithmetical work. Under these conditions there is no need to consider the reliability of the calculated value as long as it is thought of as the mean of the data from which it was computed. It is only when the calculated mean is used as an estimate of the mean of a larger population or universe that the probable error due to sampling is useful. Hence, it is not correct to write such an expression as $M = 26.4 \pm 1.5$ unless the intention is to designate the probable limits of the mean of the universe. If the calculated mean is designated, the probable error value should not appear. Confusion would be avoided by employing \tilde{M} to designate the mean of the universe as suggested on page 106.

Conditions under which probable error formulae are to be used. The conditions under which the probable error formulae are to be used have been implied in the preceding paragraphs. They are to be applied *only* when the calculated value of a statistic is taken as the estimated value of the statistic for a larger population or universe.¹ A further requirement is

¹ As pointed out on page 103, the derivation of the formulae assumes an infinite population or universe. When the random sampling has been from a finite population, Peters and Van Voorhis have shown that the formulae for the mean and the standard deviation should involve $\sqrt{1-p}$ as a factor, p being used to designate the per cent of the population included in the sample. When the population is infinite, p , of course, becomes zero. The use of the usual formulae with random samples from finite populations merely makes the probable limits conservative, which, in view of the probable presence of errors in the data, is not undesirable. Peters, C. C., and Van Voorhis, W. R. "A New Proof and Cor-

that the group of data from which calculations have been made qualify as a large random sample of this larger population. "Large" is a relative term but the formulae are generally considered to be applicable when N is not less than 30.¹ The requirement of randomness of the sample is not so simply interpreted. If the data have been selected from the universe by a process of random sampling, the requirement is satisfied, but, as pointed out on page 58, random sampling is seldom feasible in educational research. Hence, we are interested in the question, can a group of data selected in some other way qualify as a random sample?

Logically, the answer seems to be definitely in the negative. If this position is accepted, it follows that probable error formulae have little application in educational research. It may be argued, however, that the calculation of the probable error of a statistic, even when not justified on a logical basis, operates to make one conservative in generalizing from the calculated values of statistics, especially small differences. The answer to this argument is that it represents a half truth. The calculation of a probable error appears to have caused many persons to believe that they have secured a measure of the probable degree of the inaccuracy of the calculated statistic. This inference is absurd. The accuracy of a statistic is affected by other data faults.² As a practical procedure the probable error may be calculated even though the sample is not known to be random, provided it is not obviously biased, but caution should be exercised in its interpretation. As a general principle one should

rected Formulae for the Standard Error of a Mean and of a Standard Deviation," *Journal of Educational Psychology*, 24: 620-33, November, 1933.

¹ When N is less than 30 a modified technique may be used. For example, see Fisher's method of determining the significance of coefficients of correlation obtained from small samples.

Fisher, R. A. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925, pp. 159-62.

For a simplification of Fisher's method see Ezekiel, Mordecai. *Methods of Correlation Analysis*. New York: John Wiley and Sons, Inc., 1930, pp. 256-57. See also page 121.

² Data faults and their effects upon statistics are considered in the following chapter.

keep in mind that the probable error formulae are designed to give only an estimate of the effect of chance and of variable errors of measurement when the calculated value of a statistic is taken as the estimated value of the statistic for the universe. They do not yield any measure of the effect of other data faults which may be relatively large. If a sample is judged to be representative or approximately so, the calculated value of the statistic may be used as an estimate of its value for the universe, but a small probable error is not proof of a high degree of representativeness.

The value of r for a specified population, given the value of r for a non-random sample. Not infrequently the data available for calculating the coefficient of correlation obviously form a selected sample of the population for which the value of r is desired. A special case is when a coefficient of reliability for a test has been calculated from scores obtained from a single grade group and the coefficient of reliability is desired for a population including a sequence of grade groups. Kelley¹ has devised a formula for this case by assuming that the variability ($\sigma_{1.\infty}$) of the variable errors of measurement² is the same for the narrow range of talent as it is for the wide range. Using small letters to designate the statistics for the narrow range and capital letters for those of the wide range, this assumption gives

$$\sigma_{1.\infty} = \sigma_1 \sqrt{1 - r_{1I}} = \Sigma_{1.\infty} = \Sigma_1 \sqrt{1 - R_{1I}}$$

$$\frac{\sigma_1}{\Sigma_1} = \frac{\sqrt{1 - R_{1I}}}{\sqrt{1 - r_{1I}}}$$

Solving for R_{1I} we have

$$R_{1I} = \frac{\Sigma_1^2 - \sigma_1^2(1 - r_{1I})}{\Sigma_1^2}$$

This formula also involves the assumption that variable errors on the two testings to determine reliability are uncorrelated.

¹ Kelley, T. L. "The Reliability of Test Scores," *Journal of Educational Research*, 3: 370-79, May, 1921.

² For an explanation of the expression $\sigma_{1.\infty} = \sigma_1 \sqrt{1 - r_{1I}}$ see pages 132 f.

Holzinger has called attention to the doubtful validity of this assumption.¹ Hence, the formula should be used with caution.

For the correlation between true measures of intelligence and true measures of achievement in a population made up of several grade groups, Kelley² has developed a formula that may be written in the same form.

$$R_{AI} = \frac{\Sigma_A^2 - \sigma_A^2(1 - r_{AI})}{\Sigma_A^2}$$

The subscript A designates achievement and I , intelligence. The derivation involves the assumptions that σ_A , σ_I , and r_{AI} are the same for the several grade groups entering into the wide range population and that the difference between the means of the successive grade groups is constant.

Formulae have been developed for certain other cases of selection,³ but the assumptions on which they are based restrict their application. The formulae may be used in making estimates of the correlation for the desired population when the assumptions are not fully satisfied, but the results should not be considered precise.⁴

The value of r for the corresponding homogeneous population. A population is heterogeneous with reference to a given trait when it exhibits individual differences relative to it. For example, a typical fifth-grade group is heterogeneous with reference

¹ Holzinger, K. J. *Statistical Methods in Education*. Boston: Ginn and Company, 1928, pp. 251-54.

See also Brown, William, and Thomson, G. H. *The Essentials of Mental Measurement*. Cambridge: University Press, 1921, pp. 158 f.

² Kelley, T. L. *Interpretation of Educational Measurements*. Yonkers-on-Hudson: World Book Company, 1927, p. 202.

³ Pearson, Karl. "On the Influences of Double Selection on the Variation and Correlation of Two Characters," *Biometrika*, 6: 111-12, 1908.

Pearson, Karl. "On the General Theory of the Influence of Selection on Correlation and Variation," *Biometrika*, 8: 437-43, 1912.

Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, pp. 224-25.

Beatley, Bancroft. "Achievement in the Junior High School," *Harvard Studies in Education*, Vol. 18. Cambridge: Harvard University Press, 1932, p. 43.

⁴ For an illustration of the application of one of these formulae see Monroe, Walter S., and Stuit, Dewey B. "The Interpretation of the Coefficient of Correlation," *Journal of Experimental Education*, 1: 194f., March, 1933.

to mental age. If the coefficient of correlation between two measures is known for a population that is heterogeneous with reference to a third measure, the value of r for the corresponding homogeneous population can, under certain conditions, be calculated by employing the technique of partial correlation. See Chapter XI, pages 377 f.

F. INTERPRETATION OF STATISTICS

Interpretation of the mean and median. Both the mean and the median are easy to understand. The mean is simply the arithmetical average which is defined as the quotient of the sum of the several measures divided by the number of items. The calculation from a frequency distribution is simply an economical technique. The meaning of the median is expressed in its definition.¹ Both statistics designate a point on the scale of measurement. In the calculation of both the mean and the median no assumption is made with reference to the shape of the distribution, but we tend to associate these statistics, especially the mean, with the normal distribution. Hence, it is desirable to keep in mind that the mean is affected by the magnitude of each measure. An illustration of the mean is afforded by the position of a fulcrum of a balanced beam on which numerous weights have been placed. A small weight at one end will balance a larger one on the opposite side near the fulcrum.

The relative merits of the mean and the median. Since the mean is seldom equal to the median unless the distribution is perfectly normal, the question of the relative merits of these two central tendencies arises.² Unless the skewness of the distribution is marked, the mean is generally recommended. It is rigidly defined mathematically and hence lends itself to algebraic treatment. Its computation is usually a first step in the

¹ See page 72.

² For a more extended discussion of the comparative advantages and disadvantages of the different measures of central tendency and of variability see Odell, C. W. *Educational Statistics*. New York: The Century Company, 1925, pp. 104-09, 140-42.

calculation of the standard deviation, median deviation, and coefficient of correlation. The mean is based on all of the measures and on their exact magnitude. The median is easier to compute but it is less rigidly defined mathematically than the mean, and does not lend itself to algebraic treatment. The median should be used in preference to the mean when the distribution includes extreme measures that make the mean materially different from the median, when the exact magnitude of some of the measures is not known, or when ease and rapidity of computation is an important consideration.

The meaning of a measure of variability. A central tendency is a *point*; a measure of variability is a *distance*. If a normal distribution is assumed, the distance may be defined in terms of the per cent of the distribution (area under the frequency curve) that is marked off when the distance is measured in each direction from the central tendency and perpendiculars are drawn at the points thus located. In the case of the standard deviation (σ) this per cent is 68.27. For the median deviation ¹ the per cent is 50.

When comparing standard deviations or median deviations, the absolute magnitude of the measures of variability may not be as significant as ratios formed by dividing each standard deviation or median deviation by the mean or median of its distribution. The quotient formed by dividing the standard deviation by the mean ² is commonly referred to as the *coefficient of variability*. It should, however, be used with caution, because the zero points for much of our educational data are arbitrary. For further reference to this topic see Chapter VIII.

The interpretation of a coefficient of correlation—an introductory statement. Both the median and the mean are easy to comprehend. The concepts that they represent are simple and the process of calculation can be rationalized. Even in the case of the standard deviation (σ), the general plan of calculation is

¹ The median deviation of a distribution of measures should be called the median deviation (*MdD*), and the term probable error (*PE*) should be used only to refer to the median deviation of a distribution of errors.

² Usually the quotient is multiplied by 100.

easily understood and the result has a graphical interpretation. The calculation of a coefficient of correlation is more complicated and hence difficult to rationalize. Furthermore, if one succeeds in rationalizing the process of calculation, he will find that he has made little progress toward the interpretation of this statistic. A coefficient of correlation is a pure number. It does not represent a point or a distance.

When one consults texts on educational statistics concerning the meaning to be associated with a given coefficient, he is told that the value of a coefficient of correlation cannot be greater than $+1.00$ or less than -1.00 ; that a positive coefficient is evidence that the larger magnitudes in one set of data tend to be paired with the larger in the other and likewise the smaller magnitudes in one set of data tend to be paired with the smaller in the other; that a negative coefficient is evidence of inverse pairing, the larger magnitudes in one set tending to be paired with the smaller ones in the other; that the magnitude of the coefficient is indicative of the completeness of this pairing, being complete when $r = \pm 1.00$; and that when the coefficient is 0.00 the pairing is on the basis of chance and no relation exists between the two sets of measures. Obviously there is some type of correspondence between the magnitude of the coefficient and the degree of relationship between the two sets of paired measures or the variables represented by them, but educational statisticians have given only superficial attention to the degree of relationship to be associated with particular numerical values of r , such as .18, .30, .50, or .75.

In 1917 Rugg¹ suggested the following general interpretations: r less than .15 to .20, correlation "negligible" or "indifferent"; r from .15 or .20 to .35 or .40, correlation "present but low"; r from .35 or .40 to .50 or .60, correlation "marked"; r above .60 or .70, correlation "high." Such interpretations are limited in meaning and they may be misleading if not actually erroneous because "negligible," "marked," "high,"

¹ Rugg, H. O. *Statistical Methods*. Boston: Houghton Mifflin Company, 1917, p. 256.

and the like are subject to two meanings. They may be used in an absolute sense, i.e., to designate degrees of relationship between zero correlation and perfect correlation, but they are more commonly used to express comparisons within a similar population. For example, a "high" man designates one that is tall in comparison with other men although his height is actually much less than that of a "high" tree or building. Hence, the interpretation of a given coefficient of correlation as "high" may be understood to mean either that the coefficient of correlation is numerically greater than most of the coefficients obtained in the field of education or that it is large in comparison with those obtained from similar populations of data. If the second meaning is intended, a general scheme of interpretation such as that proposed by Rugg cannot be used.

Coefficients calculated from the scores obtained from the administration of two forms of a test are, other things being equal, higher than those calculated from intelligence test scores and measures of silent reading ability; these are higher than those summarizing the relationship between high school marks and those received in college; and these in turn are higher than the coefficients for IQ's and measures of teaching success. Hence, a coefficient of .50 would be *very high* for IQ's and measures of teaching success, *slightly below average* for high school marks and college marks, *low* for the relation between intelligence test scores and measures of silent reading ability, and *very low* for the reliability of a test. The interpretation of "marked" suggested by Rugg would not be a satisfactory description of the relative degree of relationship in any of the four cases. This condition suggests interpretation schemes based upon the ranges of reported coefficients for different types of data.¹ But the situation is further complicated by the fact that the magnitude of a coefficient of correlation is affected by the range of the population from which the data were obtained. For example, if the population is selected from a single school

¹ The ranges of reported coefficients for several types of data are given by Odell, C. W. *Educational Statistics*. New York: The Century Company, 1925, p. 173.

grade, the coefficient of correlation between mental age and achievement will be less than if it were calculated from a population representing a sequence of two or more grades. In other words, for measures of two traits the greater the standard deviations of the distributions of these measures, the larger the calculated coefficient of correlation.¹ Hence, a comparison of coefficients of correlation, even when they have been obtained from measures of the same traits, is likely to be misleading unless the standard deviations of the distributions of the measures are taken into account.

The use of such terms as "low," "marked," and "high," is not recommended. When a writer or speaker does employ them, he should make clear whether the description of the degree of relationship is being made on an absolute or relative basis. More satisfactory interpretation procedures will be described in connection with the consideration of the uses of correlation.

The uses of correlation. The uses of correlation are indicated by the questions with reference to which coefficients may be interpreted.

1. Given the coefficient of correlation for a random sample, is there any relationship between the two sets of paired measures in the universe?
2. When r is a coefficient of reliability, what is the magnitude of the variable errors of measurement?
3. When a regression equation derived from the correlation table is used as a formula of prediction, how accurate are the predictions?
4. Given the coefficient of correlation between two sets of paired measures, what is the degree of relationship between these sets of measures, or the traits underlying them, within the population represented by them?

Only the first of these questions will be considered here. The second is treated in Chapter V, the third in Chapter X, and the fourth in Chapter XI.

The statistical significance of a coefficient of correlation. The statistical significance of a coefficient of correlation relates to

¹ It is, of course, necessary that other conditions remain the same.

the first of the questions listed above. A value of r not equal to 0.00 is evidence of the existence of a relationship between the two sets of paired measures within the population to which the coefficient applies. Of course, a very small value of r such as .05 or even .10 indicates only a very slight degree of relationship, and, if the number of cases is less than 30 or the assumptions underlying the calculation of the coefficient are not known to be fully satisfied, this interpretation should be made with caution.

A positive coefficient denotes a positive or direct relationship between the two sets of paired measures. A negative coefficient denotes an inverse relationship. Hence, if the value of r for the universe should not have the same sign as the calculated value, the meaning would be reversed. A calculated coefficient of correlation is said to be *statistically significant* when the probability is very slight that the coefficient for the universe would be zero or have the opposite sign. When the data, from which r has been calculated, have been obtained by a process of random sampling or may be assumed to qualify as a random sample, this probability may be obtained by comparing the calculated coefficient with its probable error.¹ It is a common practice to designate a coefficient as statistically significant when it is greater than four or five times its probable error. A statistically significant coefficient of correlation may be interpreted as meaning that in the universe there is at least a slight degree of relationship between the two traits or phenomena. It should be noted, however, that the procedure for determining statistical significance requires that the measures from which r has been calculated be a *random* sample of the universe. Hence, determinations of statistical significance should be made with caution. Unless the sample is large and the assumption of its random character appears justified, the label of "statistically significant" should not be applied when the coefficient is only four or five times its probable error. Many writers have judged a coefficient to be statistically significant when a critical exam-

¹ See page 104 for the formula.

ination of the evidence would reveal that the sample probably was not random. A further reason for caution in pronouncing a coefficient of correlation statistically significant is found in the fact that the calculated value may be materially affected by the choice of the intervals in the correlation table.¹ If the data from which the calculation is made involves variable errors, the value obtained will be attenuated.² Hence, the calculated value may not be the correct value for the label expressed or implied and when this is the case the usual method of determining statistical significance is not sound.

BIBLIOGRAPHY

BRINTON, W. C. *Graphic Methods for Presenting Facts*. New York: The Engineering Magazine Company, 1914. 371 pp.

A comprehensive treatise on graphic methods.

CAMP, B. H. *The Mathematical Part of Elementary Statistics*. Boston: D. C. Heath and Company, 1931. 409 pp.

An advanced text emphasizing the theoretical phases.

CHADDOCK, R. E. *Principles and Methods of Statistics*. Boston: Houghton Mifflin Company, 1925. 471 pp.

This text is written from the standpoint of economics and the closely related social sciences. It contains, however, much that is useful to the educational research worker. Chapters are given to index numbers and time series and a brief description is given of the use of Hollerith equipment. The discussion of sampling and probable errors is especially good.

DUNLAP, J. W., and KURTZ, A. K. *Handbook of Statistical Nomographs, Tables, and Formulas*. Yonkers-on-Hudson, New York: World Book Company, 1932. 163 pp.

See pages 64-65 for a brief description of the contents of this handbook.

EZEKIEL, MORDECAI. *Methods of Correlation Analysis*. New York: John Wiley and Sons, 1930. 427 pp.

A comprehensive text on the correlation techniques. Written from the standpoint of agricultural economics, but containing much that is applicable to educational problems where correlation analysis is appropriate. The treatment is "advanced," but written with unusual clarity.

¹ See page 102.

² See page 151.

FISHER, IRVING. *The Making of Index Numbers*. Boston: Houghton Mifflin Company, 1922. 526 pp.

A comprehensive text on index numbers by the leading authority in the field.

FISHER, R. A. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1928. 269 pp.

An advanced text in statistical method which contains chapters on tests of goodness of fit, independence, and homogeneity; tests of the significance of means, differences of means, and regression coefficients; intraclass correlations and the analysis of variance; and principles of statistical estimation.

GARRETT, H. E. *Statistics in Psychology and Education*. New York: Longmans, Green and Company, 1926. 317 pp.

A text for the student without extensive mathematical training.

HOLZINGER, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928. 372 pp.

Somewhat more advanced than the texts by Garrett and Odell.

JONES, D. C. *A First Course in Statistics*. London: G. Bell and Sons, 1924. 301 pp.

An advanced text which presents an extended discussion of frequency curve-fitting.

KARSTEN, K. G. *Charts and Graphs*. New York: Prentice-Hall, Inc., 1923. 724 pp.

A comprehensive text on graphic methods.

KELLEY, T. L. *Statistical Method*. New York: The Macmillan Company, 1923. 390 pp.

An advanced text widely used in educational research.

MILLS, F. C. *Statistical Methods: Applied to Economics and Business*. New York: Henry Holt and Company, 1924. 604 pp.

Contains chapters on index numbers, analysis of time series, and method of least squares.

ODELL, C. W. *Statistical Method in Education*. New York: D. Appleton-Century Company, 1935. 457 pp.

A text for the student without extensive mathematical training.

OTIS, A. S. *Statistical Method in Educational Measurement*. Yonkers-on-Hudson, New York: World Book Company, 1925. 337 pp.

Contains an extended treatment of percentile curves and information on the applications of statistical techniques to educational measurements.

120 STUDY OF EDUCATIONAL PROBLEMS

RIETZ, H. L. *Mathematical Statistics*. Chicago: Open Court Publishing Company, 1927. 181 pp.

A brief, but advanced text in statistical method which contains derivations of a number of important formulae.

RIETZ, H. L., et al. *Handbook of Mathematical Statistics*. Boston: Houghton Mifflin Company, 1924. 221 pp.

An important reference book on general statistics.

WALKER, H. M. *Studies in the History of Statistical Method with Special Reference to Certain Educational Problems*. Baltimore: Williams and Wilkins Company, 1929. 230 pp.

WALKER, H. M. *Mathematics Essential for Elementary Statistics*. New York: Henry Holt and Company, 1934. 246 pp.

A valuable source for a number of details which are not usually treated in texts on statistical method.

WILLIAMS, J. H. *Graphic Methods in Education*. Boston: Houghton Mifflin Company, 1924. 319 pp.

YULE, G. U. *An Introduction to the Theory of Statistics*. London: Charles Griffin and Company, 1927. 422 pp. (Eighth edition).

An excellent advanced text.

CHAPTER V

THE FAULTS OF DATA AND THEIR EFFECTS

Purpose of this chapter. The calculated value of a statistic may not be the correct value even when no arithmetical errors have been made in the process. For example, the mean calculated from a frequency distribution may not be the correct value of the mean of the data. On page 103 it was pointed out that frequently the value of the mean or other statistic calculated from a sample of data is labeled or used as the value of the statistic for a larger population or universe. When this is done, the "calculated value" is likely not to be precisely the "correct value" for this label. Other types of labels are explicitly or implicitly assigned to calculated or obtained values of statistics. A term is needed to refer to the degree of agreement between the "correct value" of the statistic as labeled and the "obtained value." The word accuracy should not be used because it is widely employed with a different meaning. A term with a broader meaning is needed. "Dependability" is suggested.

The dependability of the calculated value of a statistic will vary with the label explicitly attached to it or implied in its interpretation. The types of labels may be illustrated by considering those that may be attached to the mean calculated from a frequency distribution of achievement test scores. These labels are

1. Mean of the obtained scores.
2. Mean of the true scores of the group tested under standard testing conditions.
3. Mean of the true scores of a larger population or universe.
4. Mean of true measures of achievement as specified of the population tested.
5. Mean of true measures of the achievement as specified of a larger population or universe.

The value of the mean calculated from the frequency distribution *may be* 100 per cent dependable when assigned the first label, but this is not necessarily the case. In general, the dependability of a calculated value will decrease from label to label in the order given.

It is the purpose of this chapter to consider the causes that contribute to undependability and the effects of these causes, commonly referred to as "data faults," upon the various statistics. In certain cases a more correct value of the statistic as labeled may be calculated. In other cases the probable limits of the correct value may be determined. In general, however, the degree of dependability can only be estimated. Hence, it is important that an investigator understand the possible faults of data and their effect upon the various statistics. Estimating the dependability of the findings of educational research is the first step in their interpretation.

In the following pages arithmetical accuracy of the calculations will be assumed. It will also be assumed that the number of decimal places or significant figures is consistent with the approximate nature of the data.¹ Attention will not be given to faults of procedure or to the erroneous interpretation of statistics. The restriction of the scope of this chapter to the faults of data and their effects is not intended to imply that other matters are not important. As a matter of fact, many published studies are to be criticized for other reasons. Some are not wisely planned. The investigator may have failed to gather all of the data his problem called for. He may have failed to use appropriate statistical techniques. He may have allowed his prejudice to influence his interpretation of his findings.

A. TYPES OF DATA FAULTS

The faults of quantitative data. When statistics are interpreted without generalization, two general types of data faults require consideration—(1) errors, and (2) failure of the data as a group to conform to the conditions (assumptions) underlying

¹ See pages 62 f.

ing the formulae used or associated with the interpretation of a mean, standard deviation, coefficient of correlation, or other statistic. When the interpretation of a statistic is extended to a larger population or universe, the determination of the dependability of the calculated value requires also consideration of the degree to which the data collected are representative of the population for which the generalization is desired.

In the following pages these faults are first described briefly. Then each is considered in detail, giving attention to their causes and their effects upon statistics.

Errors. The difference between a measure and the criterion by which it is judged is called an *error*. Suppose the height of a child is measured as 57.5 inches; the error is .5 if 57.0 inches is the criterion; -1.5 if 59.0 inches is the criterion, and so on. The criterion by which an obtained measure is to be judged is implied in the label given it or a derived statistic when that label is precisely expressed. In the field of physical measurements, little attention is given to this principle because the criterion is usually the "true measure." But, in the case of test scores and certain other types of educational data, we may employ any one of several labels, each implying a different criterion. The precise meaning of the label "obtained score" varies with the conditions under which the test was administered. First-trial scores and second-trial scores indicate two variations. "True score," which has a meaning corresponding to "true measures" in the physical realm, usually implies "standard testing conditions." In addition, we have labels specifying measures of a designated ability or trait.¹ Considering the variations possible, the necessity of precise labels is obvious. Unfortunately, complete and precise labels are seldom attached to educational data. The term "score" is frequently used in the sense of "true score." Sometimes "score" is apparently used with the meaning of measure of a particular ability or trait. As a basis for considering errors in data, it

¹ The reader may find it helpful to refer to the discussion of the meaning of the measurement of traits and abilities on pages 172 f.

is necessary to give careful attention to the labels attached to them.

Any obtained measure may be thought of as equivalent to the algebraic sum of the corresponding true measure and an error. If we use X to designate an obtained measure and X_{∞} the corresponding true measure, their relation may be expressed as follows:

$$X = X_{\infty} + e$$

The error designated by e may be either positive or negative. For example, if a pupil's true score is 42 and the obtained score is 37, this measure is to be thought of as involving a negative error of -5 , i.e., $37 = 42 - 5$.

In educational research a worker is usually concerned with groups of data rather than with individual items. When attempting to determine the effect of the data faults upon a statistic calculated from a group of data, it is necessary to think of the error in each item as consisting of two parts—one systematic and the other variable. In terms of symbols this relationship may be expressed as follows:

$$X = X_{\infty} + e_{\text{sys}} + e_{\text{var}}$$

As the term implies, a systematic error is one that conforms to some system. An infinite variety of systems of errors is theoretically possible, but in educational research our concern is with a systematic error that affects all measures of a given group in the same direction. For example, if four and one-half minutes are allowed in administering a speed test whose specified time limit is four minutes, and none of the pupils finish, the size of the systematic errors due to the allowance of too much time will tend to be proportional to the scores on the test, the greater the score, the greater the error. The simplest type of systematic error is one that is constant, that is, the same for all items of data in the group. When this is the case, the term *constant error* is used. It is likely that a systematic error in educational data is rarely perfectly constant, but unless there

is obviously some relationship between the error and the magnitude of the measures, it is commonly considered as constant.

Variable errors are chance errors. Within a group of measures they vary with respect to both magnitude and direction. In a large, unselected group of measures there are approximately as many negative variable errors as there are positive ones and if the variable errors in such a group could be separated out and assembled in a frequency distribution, its shape would approximate that of the normal probability curve and its mean would be zero. Variable errors occur in precise physical measurements. If the diameter of one hundred steel balls is measured to one-hundredth of a millimeter by means of a micrometer calliper, a second set of measurements is not likely to agree with the first. By repeating the measurement several times and using the mean of the results for each ball as its true measure, the variable errors in the first set of measurements may be calculated. Repeated measurement of human abilities and traits is not feasible because the act of measuring will usually introduce a practice effect which designates a positive systematic error. Consequently, variable errors in educational measurements cannot be isolated, but it is helpful to conceive of them as if they could be.

An idea of the magnitude and distribution of the variable errors that occur in educational measurements may be obtained by calculating the differences between the scores on two equivalent forms of a test when the interval between their administration has been so short that there has been little change in the ability of the pupils. Table VI gives the distribution of the differences between the scores made by a group of fifth-grade pupils on two forms of Monroe's General Survey Scale in Arithmetic. For example, one pupil made a score of 51 on the first trial and 59 on the second. The difference between the two scores is -8 . The differences given in Table VI are not the variable errors of either set of scores. Both sets of scores involve variable errors and one or both of them involves a

TABLE VI. DISTRIBUTION OF DIFFERENCES BETWEEN SCORES YIELDED BY TWO APPLICATIONS OF MONROE'S
GENERAL SURVEY SCALE IN ARITHMETIC TO A GROUP OF FIFTH-GRADE PUPILS

SCALE	DIFFERENCES	FREQUENCY OF DIFFERENCES
+22.5 to +25.49....	+23	1
+19.5 to +22.49....	+20	1
+16.5 to +19.49....		
+13.5 to +16.49....	+16	1
+10.5 to +13.49....	+11, +12, +13	3
+7.5 to +10.49....	+8, +9, +10	4
+4.5 to +7.49....	+5, +5, +5, +5, +5, +6, +6, +7, +7, +7	11
+1.5 to +4.49....	+2, +2, +2, +2, +2, +3, +3, +3, +4, +4	13
+1.5 to +1.49....	-1, -1, 0, 0, 0, 0, +1, +1, +1, +1	11
-4.5 to -1.49....	-2, -2, -3, -3, -3, -4	6
-7.5 to -4.49....	-5, -5, -5, -6, -7, -7, -7	8
-10.5 to -7.49....	-8, -8, -8, -8, -8, -8, -8, -8, -9, -9, -10, -10, -10	16
-13.5 to -10.49....	-11, -11, -11, -11, -11, -12, -12, -12, -12, -13, -13	10
-16.5 to -13.49....	-14, -14, -14	3
-19.5 to -16.49....	-17, -17	2
-22.5 to -19.49....	-21	1
-25.5 to -22.49....	-25	1

$N = 92$

Mean of the distribution = -2.11

systematic error.¹ In order to obtain the exact magnitude of the variable error of measurement for a given pupil, it would be necessary to secure his true score and to subtract it from the obtained score. This difference would be the variable error of measurement involved in his score.

Non-conformity with assumed group characteristics. Attention has been called in the preceding chapter to the fact that in making computations from frequency distributions the assumption is made that the measures are uniformly distributed within the several intervals or that the mid-point of an interval may be used as the average value of the measures within it. Linearity of relationship is assumed in the derivation of the formula for the product-moment coefficient of correlation. Other assumptions are made in connection with other formulae. Obviously, non-conformity with such assumptions is likely to introduce an error in the calculated value of a statistic. Slight departures from conformity are usually not very significant, but precision in research requires that the group characteristics of the data be considered and if the degree of non-conformity appears significant, the use and interpretation of the calculated values of the statistics should be limited accordingly.

In our use and interpretation of statistics we commonly assume certain conditions with respect to the data they have been derived from. For example, we commonly assume that a mean has been calculated from a distribution of measures that is approximately normal. Failure to recognize that the data may not satisfy the assumed conditions is likely to lead to erroneous interpretations.

Non-representativeness of data. In Chapter III it was pointed out that frequently it is not possible or at least not convenient to collect all of the data called for by the problem. This is almost certain to be the case whenever the conclusion desired is a generalization. When only a sample of the data

¹ The presence of a systematic error in one or both sets of scores is shown by the mean of the differences. If there were no systematic error, this mean would be approximately zero.

has been collected, any non-representativeness of this sample is a fault that must be considered. The value of a statistic calculated from a non-representative sample will usually differ from that for the total population or universe. For example, data from a bright fifth-grade class will not yield values that are the same as those for a population of typical fifth-grade pupils. If the blanks returned in a questionnaire study are from a selected group, as is not infrequently the case, the findings may not be true for the total population to which the questionnaire was mailed.

Errors of measurement and validity.¹ On page 123 the point was made that the magnitude of the error in a datum depends upon the criterion by which it is judged. The labels, which imply the criteria by which test scores are judged, may be classified under two heads (1) those that specify scores, obtained or true, resulting from the administration of the test under certain conditions, and (2) those that specify measures of a certain ability or trait. When the label is of the first type, only errors of measurement are involved. When measures of an ability or trait are specified, the possibility of errors of validity must also be considered. These classes of errors may be considered subordinate to both systematic and variable errors, thereby creating four types of errors.

Errors of measurement are those due to the variability of human responses,² the structure of the test, variations in its administration, and subjectivity in scoring pupil performances. If the label, explicit or implied, specifies "true scores for the conditions attending the administration of the test," only variable errors of measurement are involved. If the label specifies "true scores for conditions other than those that prevailed,"

¹ This discussion is in terms of test scores. With modification, it is applicable to certain other types of educational data.

² Holzinger has used the term "response errors" to designate much the same concept as is here represented by "errors of measurement." Since the "variability of human responses" is only one of several causes that are operative in the measurement process, the phrase, "errors of measurement" seems to be preferable. See Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, pp. 250-55.

a systematic error may be introduced. This will be in addition to the variable errors and an obtained score should be thought of as the algebraic sum of the corresponding true score, a variable error, and a systematic error.

Errors of validity are due to indirect measurement, i.e., the use of measures of one thing as measures of something else. As an illustration, consider the hypothetical situation in which the problem calls for measures of weight, but the investigator has no instrument for measuring the weight of children. If he substitutes measures of height for the needed measures of weight, it is obvious that variable errors are introduced.¹ These errors are not due to the process of measurement. They result from the substitution and are called variable errors of validity.

Although this illustration is an imaginary one, the conditions that it portrays are similar to those that prevail whenever the process of measurement is indirect.² For example, general intelligence, which is commonly thought of as "ability to learn," is measured indirectly by measuring certain types of achievement or what pupils have learned. Ability to spell words when writing sentences and paragraphs, as they are being composed, is measured by measuring the ability to spell dictated words. A silent reading test consisting of a series of short, unconnected paragraphs, each followed by a question to be answered before the next exercise is read, appears to measure an ability that differs significantly from the ability to read typical material. In many cases the substitution is less obvious, but when the label specifies measures of some ability or trait rather than merely test scores, the possibility of errors of validity is created.

When a substitution is made, the measures may be thought of as being in terms of a ratio or an index. Hence, the satisfactoriness of indirect or substitute measures depends upon the

¹ A coefficient of correlation between height and weight of .69 is reported by Gates.

Gates, A. I. "The Nature and Educational Significance of Physical Status and of Mental, Physiological, Social, and Emotional Maturity," *Journal of Educational Psychology*, 15: 329-58, September, 1924.

² For precise meaning of indirect measurement, see pages 144-46 and 172.

constancy of the ratio of the magnitude of the ability or trait actually measured to that of the ability or trait whose measurement is desired. Unless there is perfect correlation between true measures of the ability or trait actually measured and true measures of the specified ability or trait, variable errors of validity are introduced. In the illustration of substituting measures of height for measures of weight the satisfactoriness of using the former for the latter depends upon the constancy of the ratio of height to weight. This ratio is not the same for all children, and the variation indicates the introduction of variable errors of validity as the result of the substitution.

The presence of a systematic error of validity is a matter of concern when groups of measures are being compared or one is interested in the absolute magnitude of the measures. If the ratio of the mean of the substituted measures to the mean of the desired measures is not the same for two populations, one group of measures involves a systematic error of validity. If an unselected age group of boys is being compared with a corresponding group of girls using measures of height for measures of weight, the ratios will differ slightly for physiological reasons. Hence, a systematic error of validity would be introduced by the substitution. If two groups of girls are being compared with reference to weight using measures of height, and the members of one group were wearing high-heeled shoes and those of the other low-heeled shoes, a systematic error of validity would be introduced for this reason.

B. ERRORS OF MEASUREMENT

Causes of variable errors of measurement in test scores.¹

The responses of a pupil on successive trials of a test tend to be variable. A striking illustration of this variability has been

¹ The following references relate to variable errors of measurement:

Holzinger, K. J. "An Analysis of the Errors in Mental Measurement," *Journal of Educational Psychology*, 14: 278-88, May, 1923.

Symonds, P. M. "Factors Influencing Test Reliability," *Journal of Educational Psychology*, 19: 73-87, February, 1928.

Worcester, D. A. "Prevailing Errors in New-Type Examinations," *Journal of Educational Research*, 18: 48-52, June, 1928.

furnished by Ashbaugh¹ who arranged a fifteen-minute dictation exercise so that each test word occurred three times. The three spelling performances of about a fourth of the pupils were not consistent. Some pupils spelled a word correctly the first time and misspelled it in the later writings. Others misspelled a word the first time, spelled it correctly the second time, and misspelled it again on the third trial. Variability of performance is also shown by the fluctuations in individual learning curves. The cause of variations in the response of individual pupils in successive testings is complex. A pupil's performance at a given time is conditioned by his mental state, the effort he makes, his acquaintance with the test and with testing procedure in general, the environment under which he takes the test, the instructions given to him by the examiner, and possibly by other factors. It is possible that the ability of a pupil in a given field is subject to short time fluctuations. It is sufficient for our present purpose to note that variability of performance is characteristic of children and of adults. Some are known as erratic performers; others are classified as consistent performers.

Subjectivity in the scoring of the test papers contributes to the variable errors of measurement. Evidence of this is furnished by the numerous studies of the rating of the same examination papers by two or more persons. Although investigations of the Starch-Elliott type exaggerate the variable errors due to the subjectivity of the rating of examination papers, they are relatively large. Even when the scoring is objective, accidental errors may occur. In a recent study the Otis Self-Administering Test of Mental Ability was administered to 550 fifth-grade pupils. When the scoring was checked, it was discovered that the teachers had scored 362 papers correctly, 45 were scored one point too high, 142 were scored three points too low, and one

¹ Ashbaugh, E. J. "Variability in Spelling," *School and Society*, 9: 93-98, January 18, 1919.

Another good illustrative reference is Clark, J. R., and Vincent, E. L. "A Study of Variability in Arithmetic," *Journal of Educational Psychology*, 16: 267-74, April, 1925.

was scored four points too low. If these errors had been uncorrected, the mean point score of 29.0 would have been reduced 0.31. A trained scorer is not likely to make many errors, but if precision is desired, the work should be checked.¹

Describing the variable errors of measurement in a group of scores.² Assuming that the variable errors in a group of scores form a normal distribution whose mean is zero, their magnitude is described by a measure of the variability of this distribution. Although the variable errors cannot be isolated so that this distribution can be formulated, its standard deviation may be derived. If X_1 designates the scores yielded by a test and X_∞ the corresponding true scores, the problem is to describe the differences³ $X_1 - X_\infty$ which are the variable errors of measurement. If X is used to designate raw scores and x the corresponding deviation scores, we may write

$$X_1 = M_1 + x_1 \quad X_\infty = M_\infty + x_\infty$$

$$\text{Hence} \quad X_1 - X_\infty = x_1 - x_\infty + (M_1 - M_\infty)$$

If no systematic error is involved and N is large, M_1 approaches M_∞ and may be taken as equal to it. Hence, $x_1 - x_\infty$ may be substituted for $X_1 - X_\infty$ to represent the variable errors of measurement. The square of the standard deviation of the distribution of variable errors of measurement, the square of the standard error of measurement, is obtained by substituting in and squaring the usual standard deviation formula.

$$\begin{aligned} \sigma_{\text{var}}^2 &= \frac{\Sigma(x_1 - x_\infty)^2}{N} \\ &= \frac{\Sigma x_1^2}{N} - \frac{2\Sigma x_1 x_\infty}{N} + \frac{\Sigma x_\infty^2}{N} \\ &= \sigma_1^2 - 2r_{1\infty}\sigma_1\sigma_\infty + \sigma_\infty^2 \end{aligned}$$

¹ The following studies support this contention:

Dearborn, W. F., and Smith, W. C. "The Results of Rescoring Five Hundred Thirty Dearborn Tests," *Journal of Educational Psychology*, 20: 177-83, March, 1929.

Pintner, Rudolph. "Accuracy in Scoring Group Intelligence Tests," *Journal of Educational Psychology*, 17: 470-75, October, 1926.

² This topic is given further consideration in Chapter VII.

³ It is assumed that no systematic error is involved.

If we have a second set of measures of the ability for the same population

$$r_{1\infty} = \sqrt{r_{II}} \text{ and } \sigma_{\infty} = \sigma_1 \sqrt{r_{II}} \quad (\text{See page 151.})$$

Substituting these values ¹

$$\begin{aligned} \sigma_{e_{\text{var}}}^2 &= \sigma_1^2 - 2\sqrt{r_{II}}\sigma_1\sigma_1\sqrt{r_{II}} + \sigma_1^2 r_{II} \\ &= \sigma_1^2 - 2\sigma_1^2 r_{II} + \sigma_1^2 r_{II} \\ &= \sigma_1^2 - \sigma_1^2 r_{II} \\ &= \sigma_1^2(1 - r_{II}) \\ \sigma_{e_{\text{var}}} &= \sigma_1 \sqrt{1 - r_{II}} \end{aligned}$$

Since the median deviation (probable error) is easier to interpret we generally use the formula ²

$$PE_{e_{\text{var}}} = PE_{1.\infty} = .6745\sigma_1 \sqrt{1 - r_{II}}$$

It should be noted that r_{II} and σ_1 are derived from the same population. If σ_1 is not approximately equal to σ_I , the following formula should be used.

$$PE_{1.\infty} = .6745 \frac{\sigma_1 + \sigma_I}{2} \sqrt{1 - r_{II}}$$

This derivation of the formula for the probable error of measurement is based on two assumptions: first, that the number of cases (N) is large enough so that M_1 may be taken as equal to M_{∞} ; second, that the variable errors of measurement are uncorrelated with the true measures with which they are combined. The use of the formula introduces the assumption that the variable errors in a large unselected group of test

¹ It should be noted that in making these substitutions the assumption is introduced that the variable errors of measurement are uncorrelated with the true scores with which they are combined. (See page 205.) This assumption does not appear to be fully satisfied. Hence, the formula for $\sigma_{e_{\text{var}}}$ should be regarded as only an approximation.

² Several symbols have been used to designate "probable error of measurement." PE_{Meas} has the merit of being essentially an abbreviation, but the symbol $PE_{1.\infty}$ seems to be preferable. Similarly, $\sigma_{1.\infty}$ is recommended instead of $\sigma_{e_{\text{var}}}$.

scores and the variable errors in the scores resulting from a large number of administrations of a test to a single pupil, form similar normal distributions. Unless the structure of the test is defective, this assumption is probably adequately satisfied.

The probable error of measurement gives merely the limits between which we may expect to find 50 per cent of the variable errors of measurement of a typical group of scores. For example, if the probable error of measurement for a given test has been found to be 4.0, 50 per cent of the variable errors of measurement are greater than ± 4.0 , approximately one-half of them being positive. It also means that 50 per cent of them are not larger than ± 4.0 . In the case of a given pupil, we can state only the chances that the variable error of measurement of his score does not exceed certain limits. In the above illustration the chances are just even that the variable errors of measurement in his score is not larger than ± 4.0 . The chances are 4.6 to 1 that it is between -8.0 and $+8.0$. The chances for other limits also may be stated.¹

A probable error of measurement of 5.0 indicates relatively large variable errors when the mean score is 25.0, but relatively small ones when the mean score is 150.0. The magnitude of the mean score depends upon the size of the unit of the scale of measurement and the location of the zero point. Since the location of the zero point is usually arbitrary, the ratio of the probable error of measurement to the mean has a limited significance, and ratio comparisons of tests, with reference to the magnitude of variable errors of measurement in the scores yielded by them, should be made with caution. Comparisons in terms of age or grade units are likely to be more dependable.

Magnitude of variable errors of measurement to be expected in test scores. The variable errors of measurement in test scores are much greater than the corresponding errors in physical measurements. In a critical study of silent reading tests² it was

¹ See page 106.

² Monroe, W. S. "A Critical Study of Certain Silent Reading Tests," *University of Illinois Bulletin*, Vol. 19, No. 22, *Bureau of Educational Research Bulletin*, No. 8. Urbana: University of Illinois, 1922. 52 pp.

shown that the probable error of measurement for some tests was greater than 25 per cent of the mean score. In fact, for Brown's Silent Reading Test, it was found to be more than 50 per cent. In the tests which make up the Illinois Examination only twelve of forty-two probable errors of measurement which were calculated were greater than 10 per cent of the mean score.¹ The authors of the New Stanford Achievement Test² announce that the probable error of measurement for this battery of tests is approximately two months of educational achievement. It is likely that these authors have succeeded in reducing the variable errors of measurement to a lower minimum than has been attained by most other test makers. This has been accomplished in part through extending the length of the test.³

In another place the senior author has discussed the relative magnitude of the errors in the scores yielded by standardized tests and the errors in the marks assigned to examination papers.⁴ The evidence presented indicates that the variable errors of measurement for a number of widely used standardized educational tests are only slightly less than the variable errors of measurement for examinations of the essay type.

Causes of systematic errors of measurement in test scores.

The causes of systematic errors of measurement in test scores consist of those influences that affect all scores of a group in the same direction. They may be classified under the following heads: (1) the administration of the test including the time allowed, the directions given to the pupils, and the attitude of the examiner; (2) acquaintance of pupils with the general procedure of testing; (3) acquaintance of pupils with the type of exercises in the test; (4) attitude of pupils toward the test

¹ Monroe, W. S. "The Illinois Examination," *University of Illinois Bulletin*, Vol. 19, No. 9, *Bureau of Educational Research Bulletin*, No. 6. Urbana: University of Illinois, 1921, p. 49.

² Kelley, T. L., Ruch, G. M., and Terman, L. M. *New Stanford Achievement Test Manual of Directions*. Yonkers-on-Hudson: World Book Company, 1929. 16 pp.

³ See pages 208-09.

⁴ Monroe, W. S., and Souders, L. B. "The Present Status of Written Examinations," *University of Illinois Bulletin*, Vol. 21, No. 13, *Bureau of Educational Research Bulletin*, No. 17. Urbana: University of Illinois, 1923, pp. 30-42.

which is likely to be influenced by the manner in which the test is administered; and (5) bias of the person rating the performances when this process is subjective.

No specific procedures can be prescribed for identifying the causes that are operative in a particular case, but a few illustrations will indicate the magnitude of the systematic errors to be expected in test scores. In the following illustrations the systematic errors of measurement are referred to as constant although they may be more complex.

Illustrations of evidence of the presence of constant errors of measurement in test scores. If two forms of a test are administered to a group of pupils, the second set of scores will be subject to a "practice effect." The difference¹ between the mean of the first-trial scores and the mean of the second-trial scores is an index of the relative magnitude of the constant error in the two sets of scores. This difference, however, should not be interpreted as being the true magnitude of the total constant error of the second-trial scores. It is possible that the first-trial scores also involved a constant error due to failure to secure standard testing conditions or to other causes. However, when the means of the two sets of scores are not equal, we have evidence of the presence of a constant error in at least one set.

The relative magnitude of the constant error in second-trial scores varies, but the following cases are probably typical. The Illinois General Intelligence Scale² was given twice to several hundred pupils in Grades III to VIII inclusive. After making due allowance for the inequality of the two forms of this scale³ the difference between the means of the two sets of scores was approximately five points, or six months of mental age. In the eighth grade, in which somewhat unusual testing conditions appear to have prevailed, the difference was con-

¹ It will, of course, be necessary to make an appropriate allowance for any lack of equivalence of the two forms in comparing the means.

² Monroe, W. S. "The Illinois Examination," *University of Illinois Bulletin*, Vol. 19, No. 9, *Bureau of Educational Research Bulletin*, No. 6. Urbana: University of Illinois, 1921, p. 69.

³ *Ibid.*, p. 10.

siderably greater. For Monroe's General Survey Scale in Arithmetic administered to the same groups, the difference between the mean of the first-trial scores and that of the second-trial scores was approximately 3.2 points in Grades III to V, and 4.5 points in Grades VI to VIII. Two forms of the Thorndike-McCall Reading Scale were given to several groups of pupils. The mean of the first-trial scores (Form 2) was 47.78, the mean of the second-trial scores (Form 3), 51.69.

When any considerable period of time elapses between two trials on a given intelligence test, the instruction which pupils receive during this interim may materially influence their second trial scores. In an investigation¹ by the Bureau of Educational Research it was found that for a group of 134 children the mean increase of the second-trial scores on the Illinois General Intelligence Scale over the first-trial scores was equivalent to slightly more than four years in mental age. The two trials were six months apart and hence the normal increase in mental age to be expected would be six months. If we assume that the first-trial scores did not involve a constant error, it follows that the constant error introduced in the second-trial scores was somewhat greater than three and one-half years of mental age. Investigation revealed that the language instruction of these pupils during the period between the two testings functioned as coaching for the test.

Table VII gives certain gains in achievement which were obtained in an experiment to determine the relative effect of the number of sections into which a class was divided.² The six experimental groups were taught under conditions considered to be the same with the exception of the difference in sectioning. The tests used were Monroe's Standardized Silent Reading Test I, Revised, and Monroe's General Survey Scale in Arithmetic. Form 1 of these tests was given early in October, Form 2, the first of February, and Form 1 was again

¹ Monroe, *op. cit.*, pp. 69-70.

² Monroe, W. S. "The Constant and Variable Errors of Educational Measurements," *University of Illinois Bulletin*, Vol. 21, No. 10, *Bureau of Educational Research Bulletin*, No. 15. Urbana: University of Illinois, 1923, p. 15.

administered the following May. The first gains were calculated by subtracting the mean of the October scores from that of the February scores; the second, by subtracting the mean of the February scores from that of the May scores. The two forms of these tests have been shown to be slightly lacking in equivalence, especially in the case of reading rate.¹ The gains in Table VII, however, are evidence of the presence of systematic errors in addition to those resulting from the slight non-equivalence of the different forms.

TABLE VII. TWO SETS OF GAINS IN ACHIEVEMENT WHICH INDICATE THE PRESENCE OF CONSTANT ERRORS IN CERTAIN SETS OF SCORES, FIFTH GRADE

GROUP	NO. OF PUPILS	READING RATE		READING COMPREHENSION		ARITHMETIC	
		I	II	I	II	I	II
A	70	27.93	-15.78	.96	.35	23.82	21.45
B	72	3.67	22.11	1.21	1.86	14.72	5.44
C	326	-4.77	33.25	.92	2.06	12.07	6.36
D	133	-6.60	22.90	.82	.95	17.04	10.09
E	157	9.29	27.35	1.48	2.12	10.65	5.83
F	143	-9.26	41.48	.08	2.36	4.69	5.38

On the basis of our knowledge of the practice effect of testing we should expect the first gains to be larger than the second gains unless the variations of experimental conditions materially influenced the achievement of the pupils, which is extremely unlikely. We find in both reading rate and reading comprehension that the first gains are less than the ones for the second period except for Group A. In three cases the first gain is negative. In arithmetic the first gain is larger than the second in all cases except one. The smaller gains during the first semester than during the second, and particularly the negative gains, are indicative of the presence of a constant error in at least one of the sets of scores from which the gains were computed. The gains in reading rate shown for Group A are also

¹ Monroe, W. S. "The Illinois Examination," *University of Illinois Bulletin*, Vol. 19, No. 9, *Bureau of Educational Research Bulletin*, No. 6. Urbana: University of Illinois, 1921, pp. 12-18.

interesting—from October to February there is a very marked increase in rate; for the second semester the gain is negative. This suggests that the mean February score was too large, i.e., it involved a positive constant error.

In another investigation¹ conducted by the Bureau of Educational Research the mean increases in mental age scores during a period of six months for two groups of children, each numbering about 3000, were found to be .4 years and .9 years. During the next six months for the same two groups the increases were 1.4 years and 1.0 years respectively. The normal increase in mental age during either of these intervals is, of course, six months. The obtained increases for the first period might be expected to be somewhat greater because of the presence of a constant error introduced by the practice effect of testing. However, in one case the difference between the first- and second-trial scores is less than six months, and in both the increase is less than the corresponding differences between the second- and third-trial scores. The facts of this illustration become even more striking when we note that the total of the two gains for the first group is 1.8 years and that for the second 1.9. Thus, when the interval of twelve months is considered, the total increase in mental age score is approximately the same for the two groups. On the other hand, if the two intervals of six months are taken, the increases in mental age scores are radically different for the two groups. Although our knowledge of mental growth is limited, it does not appear possible to explain the inconsistencies noted except on the basis of a constant error in some of the sets of scores.

In each of the two preceding illustrations we have evidence of the presence of a constant error in at least some of the groups of test scores, but the cause is obscure. Furthermore, the exact magnitude of the constant error is unknown. The obscurity of the cause is due in part to the large number of teachers and

¹ Odell, C. W. "The Use of Intelligence Tests as a Basis of School Organization and Instruction," *University of Illinois Bulletin*, Vol. 20, No. 17, *Bureau of Educational Research Bulletin*, No. 12. Urbana: University of Illinois, 1922. 78 pp.

pupils participating in each of these educational experiments. The errors may have been due to changes in the interest and attitude of the teachers and pupils toward the test. However, it was not possible to secure any direct evidence on this point. The fact that the cause is obscure makes the possible presence of constant errors in such data a serious matter and tends to arouse suspicions regarding the accuracy of the measurements of achievement in large coöperative experiments.

In the illustrations cited in the preceding pages the tests used were objective, that is the test papers were scored by following specific instructions which eliminated the necessity for exercising judgment. In marking examination papers and other pupil performances where the scorer is asked to exercise judgment, much evidence has been collected to show that two persons are likely to differ widely in the scores they assign to the same pupil performances. These differences are due in part to the presence of a systematic error resulting from the fact that one of the scorers tends to be more liberal than the other. In an investigation¹ several sets of pupil performances, for which the scoring was rather highly subjective, were rated independently by two persons under the supervision of a third. A portion of one table from this report is reproduced as Table VIII to furnish evidence of the presence of a systematic error in the scores assigned by one or both of the scorers. The entries in the column headed "difference of mean scores" were obtained by subtracting the mean of the scores assigned by the second scorer from the mean of those assigned by the first scorer. Some of these differences are relatively large. It appears that a scorer is not always consistent with respect to his systematic error. Scorers Y and K show negative differences for two sets of papers and a positive difference for a third set. In the same investigation, eighty-six compositions were rated independently by two persons using the Willing Scale for Measuring Written Com-

¹ Monroe, W. S. "A Critical Study of Certain Silent Reading Tests," *University of Illinois Bulletin*, Vol. 19, No. 22, *Bureau of Educational Research Bulletin*, No. 8. Urbana: University of Illinois, 1922. 52 pp.

position. The difference between the means of their scores was 6.7.

TABLE VIII. SUBJECTIVITY OF SCORING REPRODUCTIONS BY THE WORD-COUNTING METHOD

TEST	FORM	GRADE	NUMBER OF PAPERS	SCORERS	DIFFERENCE OF MEAN SCORES
Memory	I	IV	27	Y—K	— 5.1
Memory	I	VII	123	Y—K	— 7.5
Memory	II	VII	31	Y—K	+ 4.1
Memory	II	IV	116	Y—C	— 2.0
Memory	I	IV	92	Y—C	— 9.9
Memory	II	VII	100	Y—C	— 8.2
Reproduction	I	IV	94	L—K	+ 6.8
Reproduction	II	IV	68	L—K	+ 4.7
Reproduction	II	IV	31	L—C	— 1.6
Reproduction	I	VII	117	M—F	— 0.5
Reproduction	II	VII	113	F—C	— 6.0
Brown	I	IV	111	T—My	+12.8
Brown	II	IV	110	T—My	+ 6.9
Starch (No. 7)	I	VII	119	M—C	— 5.8
Starch (No. 6)	II	VII	121	M—C	— 2.0

When a test has been standardized for age groups and the norms are used as a basis for translating the point scores into age scores, any systematic error in the norms will introduce a systematic error in the age scores. A similar statement may be made relative to intelligence quotients. Furthermore, the intelligence level indicated by a given IQ, say one of 112, depends upon the test from which it was computed.¹

The magnitude of constant errors in test scores. The preceding illustrations reveal a significant characteristic of the constant error in test scores. They conform to no law. In the case

¹ Kefauver, G. N. "Need of Equating Intelligence Quotients Obtained from Group Tests," *Journal of Educational Research*, 19: 92-101, February, 1929.

See also

Carroll, H. A., and Hollingworth, L. S. "The Systematic Error of Herring-Binet in Rating Gifted Children," *Journal of Educational Psychology*, 21: 1-11, January, 1930.

Cattell, Psyche. "Why Otis' 'I.Q.' Cannot Be Equivalent to the Stanford-Binet I.Q.," *Journal of Educational Psychology*, 22: 599-603, November, 1931.

of a given test it is not possible to make a general determination comparable to the probable error of measurement which will be applicable to all groups of scores obtained from administering it. When first-trial scores are compared with second-trial scores, the latter are usually found to involve a positive constant error, but the magnitude of the error varies and it may be negative. For certain tests Thorndike¹ has reported the effect of the practice due to the repetition of the test to be about 10 per cent of the mean magnitude, but the relative magnitude may be expected to vary with the experience of the subjects, the type of test, and its structure. Furthermore, practice is only one of several causes that contribute to the constant error in a group of test scores. Hence, in a particular case, even an experienced investigator can only estimate roughly the magnitude of the constant error in a group of test scores. This fact makes constant errors especially troublesome in educational research.

Systematic errors in other types of data. In making estimates of traits by means of rating scales, a phenomenon may occur which is called the "halo effect."² For example, let us suppose that the rater recognizes that the individual being rated is outstanding in a given trait. He may be, in the estimation of the rater, the most honest or dishonest person the rater has ever known. This high estimate or low estimate of one characteristic tends to cause systematic errors in the ratings of other traits of this individual. The rater is biased with respect to him. Somewhat the same phenomenon has been observed when the individuals rated are well known by the rater. Friendship may tend to bias the scores assigned and in some cases familiarity may breed contempt.³

¹ Thorndike, E. L. "Tests of Intelligence; Reliability, Significance, Susceptibility to Special Training and Adaptation to the General Nature of the Task," *School and Society*, 9: 189-95, February 15, 1919.

² Thorndike, E. L. "A Constant Error in Psychological Ratings," *Journal of Applied Psychology*, 4: 25-29, March, 1920.

³ See Knight, F. B. "The Effect of the 'Acquaintance Factor' upon Personal Judgments," *Journal of Educational Psychology*, 14: 129-42, March, 1923.

(Continued next page)

Persons who possess a given trait in a high degree tend to under-rate themselves, and persons who are very deficient with respect to the trait tend to over-rate themselves. Cogan, Conklin, and Hollingworth¹ have reported that individuals tend to over-rate themselves on desirable traits and under-rate themselves on undesirable traits. Remmers² has shown that distinguished students tend to under-rate themselves when their self-ratings are compared with ratings made by other persons. On the other hand, the differences between self-rating of undistinguished students and ratings of them made by other persons are evidences of variable rather than systematic errors. The student interested in the errors of self-ratings should also consult the researches of Cattell,³ Hoffman,⁴ Hurlock,⁵ Shen,⁶ Trow and Pu,⁷ and the summary of the validity of self-ratings given by Symonds.⁸

It is well known that systematic errors are a frequent limitation of questionnaire data. The wording or arrangement of the questionnaire items may cause a systematic error. For example, Mathews has reported that respondents tend to give a higher per cent of affirmative replies in a questionnaire of the multiple-response type to the items printed on the extreme left position.⁹

Shen, Eugene. "The Influence of Friendship upon Personal Ratings," *Journal of Applied Psychology*, 9: 66-68, March, 1925.

See also discussion of rating in Chapter III on pages 51-55.

¹ Cogan, L. C., Conklin, A. M., and Hollingworth, H. L. "An Experimental Study of Self-Analysis, Estimates of Associates, and the Results of Tests," *School and Society*, 2: 171-79, July 31, 1915.

² Remmers, H. H. "Distinguished Students—What They Are and Why," *Studies in Higher Education*, 15, Bulletin of Purdue University, Vol. 31, No. 2. Lafayette, Indiana: Purdue University, 1930. 36 pp.

³ Cattell, J. McK. *American Men of Science*. New York: Science Press, 1927. 1132 pp. (First edition, 1906.)

⁴ Hoffman, G. J. "An Experiment in Self Estimation," *Journal of Abnormal and Social Psychology*, 18: 43-49, April, June, 1920.

⁵ Hurlock, E. B. "A Study of Self-Ratings by Children," *Journal of Applied Psychology*, 11: 490-502, December, 1927.

⁶ Shen, Eugene. "The Validity of Self-Estimate," *Journal of Educational Psychology*, 16: 104-07, February, 1925.

⁷ Trow, W. C., and Pu, A. S. T. "Self-Ratings and the Chinese," *School and Society*, 26: 213-16, August 13, 1927.

⁸ Symonds, P. M. *Diagnosing Personality and Conduct*. New York: The Century Company, 1931, pp. 109-11.

⁹ Mathews, C. O. "The Effect of the Order of Printed Response Words on an

Systematic errors may occur in other types of educational research data. In historical research the investigator may record information which supports a point that he is seeking to establish to the neglect of evidence in opposition. In preparing a critical summary of research the investigator may neglect, or criticize unduly, research which opposes his own point of view. Although such limitations of data are somewhat different from systematic errors in test scores, it seems justifiable to apply the term to them. They represent a particularly vicious form of error since the investigator is unlikely to recognize their presence.

C. ERRORS OF VALIDITY

The cause of variable errors of validity. The general cause of errors of validity was given on page 129 as the use of substitute data or indirect measurement. Our understanding of this cause will be augmented by considering the types of substitutions that are commonly made. Whenever the data collected are not precisely those that are specified by the label attached to them or to derived statistics or implied in the interpretation, a substitution is introduced. When measures of human abilities and traits are specified or implied by the label, the following appear to be the more important types of substitution.

1. A measure of one achievement substituted for a measure of a different achievement. This case includes the substitution of a measure of an ability functioning under certain conditions for a measure of what is called the same ability functioning under different conditions.
2. A measure of a sample of the achievement substituted for a measure of the whole.
3. A measure of a combination of intelligence and acquired ability substituted for a measure of the results of instruction, usually restricted to certain instruction.
4. A measure of immediate ability substituted for a measure of the residue of ability at a later date.

The first type of substitution is very general. A test measures directly the ability to respond to the exercises of the test Interest Questionnaire," *Journal of Educational Psychology*, 20: 128-34, February, 1929.

under the conditions of its administration. It is not easy to specify this achievement. Its nature is suggested by the content of the test and the conditions attending its administration, but inconspicuous features may be influential in determining what is measured directly. The influence of subtle characteristics of test exercises is illustrated in studies of true-false tests. It has been shown that certain words and phrases are characteristic of many exercises for which the answer "false" should be made. For example, statements containing "always" or "never" are false in two out of three cases.¹ In fact, in one investigation² they were found to be false in three out of four cases. Other words and phrases tend to "determine" the response "true." Brinkemeier and Keys³ secured evidence indicating that a general characteristic called circumstantiality operates to suggest a response of "true." It appears, therefore, that in the case of "test-wise" students a true-false test is likely to measure directly a combination of factual information and shrewd inference. Hence, when scores yielded by such tests are considered as measures of dynamic or usable knowledge, the possibility of relatively large variable errors of validity is created.

Although we do not have similar critical studies for other types of tests, it is likely that the ability directly measured is determined in part by subtle characteristics of the measuring instrument employed. Hence, it is difficult, perhaps impossible, to determine with precision what a test measures directly. When there is a relatively large amount of writing, motor skill is likely to be involved.⁴

¹Weidemann, C. C. "How to Construct the True-False Examination," *Teachers College, Columbia University Contributions to Education*, No. 225. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 118 pp.

²Brinkmeier, I. H., and Ruch, G. M. "Minor Studies on Objective Examination Methods. III. Specific Determiners in True-False Statements," *Journal of Educational Research*, 22: 110-18, September, 1930.

³Brinkemeier, I. H., and Keys, Noel. "Circumstantiality as a Factor in Guessing on True-False Examinations," *Journal of Educational Psychology*, 21: 681-94, December, 1930.

⁴For an illustration of an attempt to correct obtained scores on a speed test for the effect of motor skills, see Courtis, S. A., and Thorndike, E. L. "Correction Formulae for Addition Tests," *Teachers College Record*, 21: 1-24, January, 1920.

Usually a test measures only a relatively small segment of achievement. Hence, in addition to substituting measures of one achievement for measures of another achievement, a measure of a small segment of achievement is substituted for a measure of a larger segment. Unless the sample is representative of the whole in the case of the persons taking the test, the variable errors of validity are increased by this substitution. When the achievement to be measured is defined as the result of instruction, there is a further contribution to the variable errors of validity because a test usually measures directly a combination of intelligence and acquired ability and the relative proportion of these factors varies from person to person. If the achievement to be measured is thought of as being relatively permanent, there is likely to be a still further contribution.

Causes of systematic errors of validity in test scores. When a substitution of data is made, it is not necessary that the mean of the substitute measures be numerically equal to the mean of the specified measures. In fact, in the case of test scores little effort has been made to establish a standard unit of measurement without which numerical equivalence could not be expected. Hence, when a researcher is concerned with only one set of data, it is not necessary to give attention to the question of the presence of a systematic error of validity. It is only when norms are introduced or two or more sets of measures are being compared that this question requires consideration. The substitution will not introduce a systematic error of validity if the ratio of the mean of substituted measures to the mean of the specified measures is the same for all groups of measures. For example, the substitution of scores on a mere information test for measures of the total achievement in such a subject as physics will not introduce a systematic error of validity if the ratio of the mean of the information test scores to the mean of measures of the total achievement is the same for the several populations. Hence, the causes of systematic errors of validity are to be sought in the conditions that produce fluctuations in the ratio of these means.

Systematic errors of measurement will affect this ratio, but it is also affected by variations in the relative magnitude of what the test actually measures and of the ability or trait specified. For example, under a given curriculum and plan of instruction calculation achievement in arithmetic bears a certain ratio to problem-solving achievement. Under these conditions the mean score on a calculation test may be considered a valid measure of the average problem-solving achievement. If the curriculum or general plan of instruction is modified with the result that this ratio is changed, the mean score resulting from a second administration of the calculation test will involve a constant error of validity when used as a comparable measure of the average problem-solving ability. If the curriculum or plan of instruction is different for two groups, the mean score of one group will probably involve a constant error of validity. As another illustration, suppose an information test is being used to measure the total achievement in such a subject as English literature. The relation existing between the mean score on the information test and the average total achievement on one date may be changed as the result of the instructor's emphasis upon information objectives and the efforts of the students to become able to respond to information tests. If this happens, a significant systematic error of validity will be introduced in the scores made on such a test at a later date. This systematic error will be in addition to that designated as practice effect.

The repeated administration of a particular type of test will tend to cause pupils to make the ability to respond to such tests their objective¹ and hence is likely to affect the ratio of what is measured directly to other achievement. This ratio is likely to be affected also by the amount and recency of instruction relating to the topics covered by the test. Hence, in considering the possibility of systematic errors of validity, it is necessary to note not only the curriculum and general plan of instruction but also less obvious conditions such as those just mentioned.

¹ Meyer, George. "An Experimental Study of the Old and New Types of Examination," *Journal of Educational Psychology*, 26: 30-40, January, 1935.

The measurement of variable errors of validity in test scores. If criterion measures are available,¹ the coefficient of correlation between them and the scores yielded by the test (r_{1c}) is an index of the total variable error in the test scores and the variable error in the criterion. This coefficient is equal to the product of the coefficient of correlation between true scores and true criterion measures and the square roots of their reliability coefficients.²

$$r_{1c} = r_{x\infty c\infty} \sqrt{r_{1I}} \sqrt{r_{c\infty}}$$

The true coefficient of validity, $r_{x\infty c\infty}$ is given by the formula³

$$r_{x\infty c\infty} = \frac{r_{1c}}{\sqrt{r_{1I}} \sqrt{r_{c\infty}}}$$

Although r_{1c} is commonly referred to as the coefficient of validity, it actually is the coefficient of reliability *and* validity.⁴ If the criterion is fallible, that is, involves variable errors of measurement, r_{1c} indicates a degree of validity that is too low. In such cases the following formula should be used in determining the coefficient of validity.

$$r_{1c\infty} = \frac{r_{1c}}{\sqrt{r_{c\infty}}}$$

The coefficient of validity, r_{1c} or $r_{1c\infty}$, is unsatisfactory as a measure of the variable errors of validity for the reasons given in connection with the coefficient of reliability. By an argument similar to that on pages 132-33, it is possible to derive a formula for the median deviation (probable error) of the differences $X_1 - X_{c\infty}$ but since criterion measures of known reliability are seldom obtainable, the formula does not have much application. See Chapter XI for an interpretation of the coefficient of correlation between the scores yielded by a test and criterion measures.

¹ Satisfactory criterion measures are usually difficult or impossible to obtain. See Chapter VII, pages 207-08.

² See page 209.

³ See page 151.

⁴ Validity is sometimes defined as inclusive of reliability. When this definition is accepted, this distinction is not justified.

The magnitude of systematic errors of validity. A coefficient of validity does not reveal whether a systematic error of validity is involved. It is a measure of the variable errors of validity. In fact, the term "validity," as it is commonly used, does not refer to systematic errors. There are no definite techniques for determining the magnitude of systematic errors of validity and estimates of experienced persons are not likely to be very dependable. About all that can be done is to indicate cases in which a systematic error is probable and to limit interpretations accordingly.¹

D. EFFECT OF DATA FAULTS

The effect of errors upon statistics.² Many persons appear to believe that the effect of errors in the data upon the mean, median, standard deviation, coefficient of correlation, and other statistics becomes negligible when the number of cases is sufficiently large. This is only partially true. A constant error in the original data makes the mean in error by the same amount and any increase in the number of cases does not decrease the magnitude of this effect, provided the constant error remains the same as the number of cases is increased.³ The same situation prevails for the median. On the other hand, a constant error does not affect the standard deviation and other measures of variability. Neither does it affect the coefficient of correlation. Systematic errors which are proportional to the size of the score do not affect the coefficient of correlation, but other types of systematic errors may affect it.

The variable errors in an unselected group of data tend to form a normal distribution whose mean is zero. The median

¹ For further consideration of the magnitude of systematic errors of validity, see Chapter VIII, page 149.

² A different type of treatment of this topic for certain statistics is given by Bowley, H. L. *Elements of Statistics*, Third Edition. New York: Charles Scribner's Sons, 1917, pp. 203-14.

³ Under certain conditions, systematic errors tend to become variable errors as the number of cases is increased. This probably occurs in a large coöperative investigation in which tests are administered by a number of different persons.

deviation of this distribution, properly called the probable error, may be calculated as indicated by the formula.¹

$$PE_{1.\infty} = .6745\sigma_1\sqrt{1 - r_{1I}}$$

In this formula r_{1I} is the coefficient of reliability and σ_1 is the standard deviation of the obtained measures.²

The effect of variable errors upon the mean cannot be determined. It is possible only to determine the probable limits of the effect for a given probability. The formula is

$$PE_{1.\infty M} = \frac{.6745\sigma_1\sqrt{1 - r_{1I}}}{\sqrt{N}}$$

Since the square root of N appears in the denominator, we may say that the effect of variable errors upon the mean varies inversely as the square root of the number of cases. This relationship is doubtless the basis of the belief that the effect of errors in the data tends to become negligible when the number of cases is large.

The probable error of measurement of the difference between two means is given by

$$PE_{1.\infty D} = \sqrt{PE_{1.\infty M_1}^2 + PE_{2.\infty M_2}^2 - 2r_{12}PE_{1.\infty M_1}PE_{2.\infty M_2}}$$

If the two sets of measures are uncorrelated, the product term becomes zero.

The probable error of measurement of individual gains (differences between scores on comparable forms of a test) is given by³

$$PE_{1.\infty G} = .6745\sigma_1\sqrt{2 - r_{1I} - r_{2II}}$$

¹ For the standard error of the probable error of measurement, see Kellogg, C. E., and Spence, K. W. "Note on the Standard Errors of Estimate and Measurement," *Journal of Educational Psychology*, 22: 313-15, April, 1931.

² If this standard deviation differs significantly from the standard deviations used in calculating the coefficient of reliability, then the coefficient of reliability should be corrected by means of Kelley's formula for the relation between ranges in obtained scores and reliability coefficients. See pages 110-11 or Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, p. 222.

³ For derivation of formula see

Kelley, T. L. "A New Method for Determining the Significance of Differences in Intelligence and Achievement Scores," *Journal of Educational Psychology*, 14: 321-33, September, 1923.

Holzinger, K. J. "Note on Professor Kelley's Formula for Determining the

Variable errors tend to make the standard deviation and other measures of variability larger than they would be otherwise. The relation between the obtained standard deviation and the true standard deviation is given by the following formula:

$$\sigma_{\infty} = \sigma_1 \sqrt{r_{1I}}$$

If the coefficient of reliability is known, this formula provides a means for correcting the calculated value of σ for the effect of variable errors. Since N does not appear in the formula, it follows that increasing the number of cases does not reduce the effect of the variable errors of measurement upon the standard deviation.

The presence of variable errors in the data tends to decrease the coefficient of correlation, an effect known as "attenuation." Several formulae ¹ have been employed to correct for the effect of variable errors of measurement. The following one contributed by Spearman ² is frequently used:

$$r_{\infty\omega} = \frac{r_{12}}{\sqrt{r_{1I}} \sqrt{r_{2II}}}$$

The use of the formula may be illustrated as follows:

$$\begin{array}{ll} r_{12} = .59 & r_{\infty\omega} = \frac{.59}{\sqrt{.85} \sqrt{.83}} \\ r_{1I} = .85 & \\ r_{2II} = .83 & r_{\infty\omega} = .70 \end{array}$$

The derivation of this formula assumes that the variable errors in the two sets of measures are uncorrelated with each other. See "Significance of Differences," *Journal of Educational Psychology*, 16: 48-51, January, 1925.

Kelley, T. L. "Professor Kelley's Reply," *Journal of Educational Psychology*, 16: 52-55, January, 1925.

¹ For a list of the formulae used in correcting for attenuation, see Dunlap, J. W., and Kurtz, A. K. *Handbook of Statistical Nomographs, Tables, and Formulas*. Yonkers-on-Hudson: World Book Company, 1932, p. 127.

² Spearman, C. "The Proof and Measurement of Association between Two Things," *American Journal of Psychology*, 15: 90, January, 1904.

Spearman, C. "Demonstration of Formulae for True Measurement of Correlation," *American Journal of Psychology*, 18: 161, 1907.

Spearman, C. "Correlation Calculated from Faulty Data," *British Journal of Psychology*, 3: 271-95, 1910.

other and with the true measures with which they are combined. Brown and Thomson ¹ have pointed out that these assumptions are frequently not satisfied. Hence, the corrected coefficient obtained should not be thought of as a highly precise determination. Correction for attenuation by means of the above formula may be regarded as yielding a best estimate of the coefficient which would be obtained if we could calculate it from true scores.

If the coefficient of reliability is known, it is possible to calculate estimated true measures commonly called regressed measures. If \bar{X}_∞ is used to designate the estimated true measures, the formula ² is

$$\bar{X}_\infty = r_{1I}X_1 + (1 - r_{1I})M_1$$

The regressed measures are only estimates of the true measures. They involve variable errors, but they are smaller than in the original measures. Their probable error is given by the expression $.6745\sigma_1\sqrt{r_{1I} - r_{1I}^2}$ instead of $.6745\sigma_1\sqrt{1 - r_{1I}}$.

The preceding discussion has dealt with the effect of errors of measurement. The general statements relative to the effect of variable errors and the effect of constant or systematic errors are also applicable to errors of validity. In the absence of appropriate criterion measures the effect of variable errors of validity can only be estimated. It should be noted, however, that the errors to be considered in a given case depend upon the label given the calculated value or implied in its interpretation. For example, teachers' marks are known to be fallible measures of achievement, but if a coefficient of correlation calculated from such data is interpreted in terms of the relationship between two sets of marks, the only errors to be considered are the accidental ones that may have occurred in copying the marks from the school records or in handling them. If, however,

¹ Brown, William, and Thomson, G. H. *The Essentials of Mental Measurement*. Cambridge: University Press, 1921, p. 158.

² This is essentially the regression equation connecting X_1 and X_I . It is assumed that $M_1 = M_I$ and $\sigma_1 = \sigma_I$. See Kelley, T. L. *Interpretation of Educational Measurements*. Yonkers-on-Hudson: World Book Company, 1927, p. 178.

the interpretation is to be made in terms of the relationship between the achievements actually represented by the marks, their reliability must be considered, and if the interpretation is in terms of the relationship between specified achievements, both reliability and validity must be considered.

Recognizing and making allowances for non-agreement with assumptions. At several places in Chapter IV attention was called to an assumption made in the derivation of a formula or to one introduced in interpreting the statistic. When the mean or standard deviation is calculated from a frequency distribution, the assumption is made that the measures within an interval are distributed so that the mid-points of the intervals may be taken as their average value. In the calculation of the mean, the positive deviations from this assumption will frequently be approximately equal to the negative ones and hence the mean will be unaffected. Sometimes, however, non-agreement with this assumption may have a material effect upon the calculated value of the mean. In calculating the standard deviation, the deviations from the mean are squared and hence there is no neutralization of the effects of non-agreement with the assumption. In frequency distributions that are normal or approximate the normal, the tendency is for the actual mean of the measures in an interval to fall nearer the mean of the entire distribution than the mid-point of the interval. When the grouping is coarse, i.e., when the interval is relatively large, the error thus introduced is important enough to require attention if precise results are desired. Sheppard's correction formula ¹ for the standard deviation is

$$\sigma_{\text{corrected}} = \sqrt{\sigma_{\text{obtained}}^2 - \frac{i^2}{12}}$$

¹ Sheppard, W. F. "On the Calculation of the Average Square, Cube, etc., of a Large Number of Magnitudes," *Journal of the Royal Statistical Society*, 6: 698-703, 1897.

Sheppard, W. F. "On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equidistant Divisions of a Scale," *Proceedings of the London Mathematical Society*, 29: 353-80, 1898.

Pearson, Karl, et al. "On the Elementary Proof of Sheppard's Formulae for

The symbol i designates the number of units in the class interval.

Since the standard deviation is involved in the calculation of the coefficient of correlation, coarse grouping also affects the value of r obtained. Sheppard's correction for a coefficient of correlation is given by the following formula ¹

$$r_{12}^{\text{corrected}} = \frac{\Sigma x_1 x_2}{N \sqrt{\left(\sigma_1^2 - \frac{i_1^2}{12}\right) \left(\sigma_2^2 - \frac{i_2^2}{12}\right)}}$$

When the unevenness of the distribution of the measures within an interval is due to other causes, there is no general procedure for making allowance for non-agreement with the assumption of uniform distribution. If non-agreement with this assumption is suspected, distributions may be formed for two or more different choices of interval points. If the calculations from the different distributions give essentially the same results, there is probably satisfactory agreement with the assumption. If the results differ,² the investigator faces the problem of determining which result is most nearly correct. In particular cases, an alert and resourceful investigator may be able to modify the plan of calculation so as to secure a more correct result. For example, in handling distributions of teachers' salaries the points of concentration may be noted and the calculations modified accordingly.

Linearity of relationship is assumed in computing the product-moment coefficient of correlation. Identification of non-conformity with this assumption is accomplished by means of the Blakeman test described on page 102. When the relationship is not sufficiently linear, the correlation ratio (η) should be used.

Correcting Raw Moments, and on Other Allied Points," *Biometrika*, 3: 308-12, 1904.

¹ In this formula and the preceding one the σ 's are expressed in terms of the scale of measurement. If they are in terms of intervals, i_1 and i_2 become unity.

² For an illustration in the case of the coefficient of correlation, see page 102.

The calculation of other statistics or their interpretation frequently implies one or more assumptions. The reader interested in this topic may locate the discussion of these assumptions by referring to the Topical Index of this volume. Some of the more important references are: coefficient of reliability calculated by means of the Spearman-Brown formula, page 202; calculation of coefficient of reliability for a greater or less range of talent, page 110; probable error of a difference, pages 150 and 308 f.; partial correlation, pages 380 f.

Making allowances for non-representativeness of data.

There are two general types of non-representativeness: (1) that caused by chance in random sampling; and (2) that due to systematic influence. The effect of non-representativeness of the second type can be calculated or estimated only when supplementary data are available and for such cases there is no general technique.¹ The effect of chance in random sampling was dealt with in the preceding chapter under the heading of "The probable limits of the value of a statistic for a universe when calculations have been made from a random sample," and hence the discussion here will be limited to certain supplementary points.

The designation of a group of data as a sample means that they have been taken from a larger population or universe. In the case of some types of data it is possible to conceive of two universes. In the case of test scores one consists of the scores resulting from an infinite number of administrations of the test to a given group of pupils under the same testing conditions. Repeated administration of a test under the same conditions is, of course, not possible, but such a universe may be thought of, and the scores obtained from a single administration under standard conditions may be taken as a random sample of it. The other type of universe involves an unlimited population of pupils. The probable error formulae given on pages 104-05 are for this type of universe and they are

¹ Certain techniques for use in connection with the coefficient of correlation have been described in Chapter IV, pages 110-11.

applicable only when the data qualify as a random sample of it. A measure of the effect of using a sample from a universe of the first type is obtained by means of the probable error of measurement. It may be noted that successive samples from the universe of performances are not independent but correlated. This condition is shown by coefficients of reliability.

The fact that we may conceive of a universe of performances, as well as a universe of subjects, has resulted in considerable confusion. Any group of test scores may be thought of as a sample of the hypothetical universe of performances by the subjects tested. But a group of scores is not necessarily a random sample of a specified universe of subjects and some persons have unconsciously shifted in their thinking from a universe of performances to a universe of subjects. A clear distinction should be made between sampling errors (effect of chance in random sampling from a universe of subjects) and variable errors of measurement (effect of sampling a universe of performances).

The effect of chance in random sampling is just as likely to make the calculated value of a statistic larger than the value for the universe as it is to make it smaller. Hence, it is possible only to calculate the probable limits of the value of the statistic for the universe. For this purpose the probable error formulae given in Chapter IV, pages 104-05, are commonly employed. It should be noted, however, that the formulae for the probable error of a mean, median, and standard deviation give the probable effect of chance plus the probable effect of variable errors of measurement.¹ If the data are fallible and it is desired to obtain the probable error of a statistic due to chance in random sampling alone, the following formulae must be used.²

¹ The coefficient of correlation may be corrected for the effect of variable errors and when such correction has been made, the probable error formula depends upon the attenuation formula used. See Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, pp. 210 f.

² Kelley, T. L. "Note upon Holzinger's Formula for the Probable Error," *Journal of Educational Psychology*, 14: 376-77, September, 1923.

Huffaker, C. L., and Douglass, H. R. "On the Standard Errors of the Mean

$$PE_{M_s} = .6745 \frac{\sigma_1}{\sqrt{N}} \sqrt{r_{1I}}$$

$$PE_{Md_s} = .8454 \frac{\sigma_1}{\sqrt{N}} \sqrt{r_{1I}}$$

$$PE_{\sigma_s} = .4769 \frac{\sigma_1}{\sqrt{N}} \sqrt{r_{1I}}$$

Since the effect of variable errors of measurement must also be considered, the above formulae are seldom used.

Another point to be noted is that the probable error of a statistic due to sampling has no connection with systematic errors of measurement, variable errors of validity, or systematic errors of validity. Hence, the probable error of a statistic due to the use of a random sample cannot be used as an index of the effects of these other faults of the data. In other words, when a statement is made of the probability that the value of the statistic for the universe lies within certain limits, there should be added the qualifying phrase "disregarding the effects of systematic errors of measurement, variable errors of validity, and systematic errors of validity."

A concluding statement. The first sentence of this chapter stated that the calculated value of a statistic may not be the correct value, even when no arithmetical errors have been made in the process, and the term "dependability" was introduced to designate the degree of agreement between the calculated value and the correct value of the statistic as labeled. The discussion of the chapter has dealt with the nature and effects of the various types of data faults, and it should be apparent to the reader that when the precise nature of the label is considered the value to which it is attached is very frequently not correct. In some cases the magnitude of the error is relatively large.

The facetious classification of statistics as representing the maximum degree of falsehood and deceit is indicative of the

Due to Sampling and to Measurement," *Journal of Educational Psychology*, 19: 643-49, December, 1928.

subtle and insidious nature of the causes contributing to undependability. An untrained investigator not infrequently makes unjustified interpretations because he is not aware of the undependability of his findings; and a person interested in making a case is likely to "discover" findings that appear to support his beliefs or wishes. There are no definite techniques for identifying the presence of systematic errors of either measurement or validity which are important data faults in the case of central tendencies or differences between central tendencies. The effect of variable errors of measurement upon the standard deviation or upon the coefficient of correlation may be eliminated if coefficients of reliability are known for the population represented by the data. There is no technique for correcting for the effect of variable errors of validity. The calculation of the probable or standard error due to sampling tends to give the impression of demonstrating the dependability of the findings, but it is obvious that this is not true. Hence, a person who has no intention of doing so may deceive himself or his audience in regard to the dependability of his findings. Many of the reported findings are not correct when the explicit or implied label is considered, and hence one who reads reports of studies should attempt to estimate the degree of dependability of the reported findings. This, however, is not easy to do. A critic is in a less favorable position than the one who conducted the research and unless a reader is supplied with adequate information, he can only estimate the dependability upon the basis of his general experience with data of the type involved.

CHAPTER VI

STUDYING THE PAST IN EDUCATION

Historical research in education. Until about 1910 history of education was a popular field of inquiry and although during recent years it has been overshadowed by the quantity production of educational research in other fields, studying the past in education still commands the attention of a number of persons. Investigating the past in education is a fascinating activity. To reveal a hitherto unknown origin of an idea or practice contemporarily regarded as new is as satisfying to the historical research worker as the discovery of a new organic compound to the research chemist. To show that a goal formulated for the education of the present is sanctioned by centuries-old concepts of educational values is both satisfying and practical. An aim or practice that is old is not necessarily worthy of continued acceptance or use, but the test of years of thought and trial cannot be ignored. Historical studies frequently contribute to the understanding of present institutions and practices. Historical findings may be as significant as those resulting from surveys and experiments. It is probable that many current absurdities in education would not be extant if the educators to blame for them had heeded the lessons of the past. Furthermore, historical research may be a source of inspiration to those engaged in educational activities.

Historical problems in the field of education. The reader who is not acquainted with this field of research can gain some idea of the problems studied and the scope of the field by noting the titles of the references in the illustrative bibliography at the end of the chapter. Although no precise classification is possible, certain types of problems may be mentioned. Some studies are restricted to a historical account of a particular per-

son or particular educational institution. A problem of this type is perhaps the simplest and the success of the investigator depends primarily upon locating sufficient authentic sources. The report is largely a statement of the facts discovered in examining the sources. A second type of problem calls for tracing the development of education or a phase of it within a specified geographical area. If this area is large and the development varied in different sections, the investigator is likely to encounter difficulty in generalizing. It is not easy to generalize in surveys of present practices over a wide area. It is more difficult to do so in a historical study. A third type of problem calls for tracing the origin or development of a movement.¹ The inadequacy of available records frequently makes the determination of origins difficult and the ramifications of a movement are not easy to identify.

Historical problems in the field of education are more varied than these three types may indicate, but they are probably sufficient to suggest the general nature of historical research in this field. A problem of the first type does not call for extensive training. After adequate sources are located, the principal requirement is systematic and careful work. For the other types of problems, a background acquaintance with educational

¹ Modern writers on historical method appear to be divided on the question of whether or not the historical research worker should attempt to show cause and effect relationships. For example, Teggart and Croce restrict the task of the historical research worker to that of revealing the characteristics of events of the past in their sequential order, while Bernheim, Fling, Langlois, and Seignobos, and Vincent sanction the making of inferences with respect to cause and effect relationships.

Teggart, F. J. *Theory of History*. New Haven: Yale University Press, 1925, pp. 61-68.

Croce, Benedetto. *History, Its Theory and Practice*. New York: Harcourt, Brace and Company, 1921, pp. 64-82.

Bernheim, Ernst. *Lehrbuch der Historischen Methode*. Leipzig: Verlag von Duncker und Humblot, 1894, pp. 492-522. (First edition, 1889.)

Fling, F. M. *The Writing of History*. New Haven: Yale University Press, 1920, pp. 146-50.

Langlois, Ch. V., et Seignobos, Ch. *Introduction aux Études Historiques*. Paris, 1898. (English translation by G. G. Berry, Henry Holt and Company, 1925, pp. 285-95.)

Vincent, J. M. *Historical Research*. New York: Peter Smith, 1929, pp. 261-76. (Reprinted from the 1911 edition of Henry Holt and Company.)

history is essential. Much interpretation of data is involved and a person who does not have an adequate background is likely to make many errors. He will also be handicapped in locating sources.

The student contemplating a historical study as a thesis should be cautioned against a hasty selection of a problem. Accessibility of adequate sources is necessary. Hence, it is unwise for a student to commit himself to a problem until examination of the accessible source materials indicates with reasonable certainty that with sustained effort and painstaking work he will ultimately produce a satisfactory report. In the general discussion of the definition of research problems in Chapter II, the necessity of a clearly defined problem as a guide for collecting data was emphasized. The definition of a problem in historical research is not so essential, but it should be sufficiently limited so that intensive study of it is possible. A defined problem will aid one in locating sources and in most cases it is necessary as a guide in selecting data from sources. A well defined problem facilitates the organization of the report and most readers desire a definition as a means of orientating their study of it.

The procedures of historical research.¹ The four phases of educational research—defining the problem, collecting data, handling them, and interpreting the findings—appear in historical inquiry but obviously in a somewhat specialized form. In the case of problems relating to the remote past, the sources of information are frequently limited and the problem must be defined to fit the available data. When the sources are more ample, definition of the problem takes the form of limiting the scope of the proposed study. Frequently it is not possible to define the problem by analyzing it into a series of subordinate questions, at least not until a preliminary survey of the available sources has been made. Collecting data is a conspicuous

¹ For a more extended account of this topic see Good, H. G. "Historical Research in Education," *Educational Research Bulletin*, 9: 7-18, 39-47, 74-78, January 8, January 22, February 5, 1930.

phase of historical research and will be treated in the following pages. Handling the data is accomplished by organizing them for effective presentation. In interpreting the data it is necessary to give attention to their accuracy and validity.

The sources of historical educational research data. In historical research a distinction is made between primary and secondary sources of data. The former comprise remains or relics and written documents which have survived from the period being studied and which represent or contain first-hand information relevant to the problem of the investigation. Written documents are of many kinds. Charters, town records, court records, constitutions, records of legislation, and reports of public officers may be classified as official documents. Unofficial documents include newspapers, magazines, letters, diaries, autobiographies, and chronicles. College and university catalogs and registers, and courses of study and syllabi of other public and private schools possibly constitute semi-official sources since they lack the legal and governmental characteristics of truly official documents and are more impersonal than documents correctly termed unofficial. Textbooks of the past, a very valuable source of historical educational research data, may be classified as relics. The title page and the preface may warrant classification as a document.

Secondary sources are accounts of the past written by persons who had access either directly, or indirectly, to primary or original sources. Typical secondary sources are reports of previous studies in the field of the problem and texts on the history of education. It may be contended that in order for an investigation to qualify as historical research some use must be made of primary sources. If the investigator does not collect at least a portion of his data from primary sources he cannot make an "original" contribution. Secondary sources may be useful in making evaluations of the appropriateness, validity, accuracy, and adequacy of the data secured from primary sources and in making inferences with respect to cause and effect relationships, where the data obtained from primary

sources by the investigator are insufficient for this purpose.

Locating sources and copying data from them. Many of the sources in a library may be located by referring to the card catalog, but if access to the stacks is permitted, it is advantageous to examine the volumes in the sections relating to the problem. Frequently, the investigator must seek sources outside the library of his own institution. Many of the larger libraries have special collections of historical documents and the investigator should seek information concerning those that relate to his problem. A graduate student will usually be able to secure this information through his adviser, but it may be necessary to make direct inquiry. For some problems, records may be sought in public offices and valuable source material, especially correspondence, may be located in the possession of private individuals. Sometimes valuable material may be located in second-hand bookstores and even in unexpected places. As the investigator reads the sources that he has located, he should be alert in noting references to additional sources. The quality of historical research is conditioned by the adequacy of original sources examined and, hence, a study should not be undertaken unless it appears feasible to obtain access to adequate sources.

In copying data from sources, the procedure is that of recording a reference to the source and the items of information relevant to the problem. A technique relative to this procedure recommended by historians is that of recording each separate item of information on a separate card or sheet with a citation to its source. For convenience the sources may be numbered and the citation may be made by writing only the number of the source and the page from which the item is taken. The arrangement of the various items of information on separate cards or sheets greatly facilitates their organization.¹ Checking

¹ The following reference may be consulted for further details of note taking:

Dow, E. W. *Principles of a Note System for Historical Studies*. New York: Century Company, 1924. 124 pp.

all items with respect to accuracy in copying is an obviously essential procedure. In selecting the items for copying, the investigator should be guided by his problem. This is especially important in the case of problems relating to the comparatively recent past for which extensive source materials are available. The investigator should also attempt to determine the accuracy and validity of the items copied.

Organizing historical data. The organization of historical data varies with the type of problem. Sometimes a chronological organization is desirable. In other cases the organization is about certain sub-topics. The purpose is to bring together related facts so that they may be effectively presented. This requires the formulation of some plan which should be made apparent to the reader. A mere recital of factual statements makes monotonous reading and a reader's estimate of a report is likely to be influenced by the author's organization of his material.

Determining the accuracy and validity of historical data.¹ Dates and other factual statements found in sources may not be accurate. The techniques described in Chapter V for identifying errors of measurement are not applicable to historical data, but the historical investigator should endeavor to determine the dependability of his sources and consequently the accuracy

¹ For further discussion of the techniques of historical criticism, the following references may be consulted:

Bernheim, *op. cit.*, pp. 236-438.

Fling, *op. cit.*, pp. 48-125.

Freeman, E. A. *Methods of Historical Study*. London: Macmillan and Company, 1886. 335 pp.

George, H. B. *Historical Evidence*. Oxford: Clarendon Press, 1909. 223 pp.

Good, H. G. "Historical Research in Education," *Educational Research Bulletin* (Ohio State University), 9: 7-18, 39-47, 74-78, January and February, 1930.

Johnson, Allen. *The Historian and Historical Evidence*. New York: Charles Scribner's Sons, 1926. 179 pp.

Langlois et Seignobos, *op. cit.*, pp. 71-208.

Marshall, R. L. *The Historical Criticism of Documents*. New York: The Macmillan Company, 1920. 62 pp.

Seignobos, Ch. *La Méthode Historique Appliquée aux Sciences Sociales*. Paris: Felix Alcan, 1909, pp. 29-93. (Second edition.)

Vincent, *op. cit.*, pp. 19-260.

of the data obtained from them. As a means of accomplishing this, he should inquire concerning the author of the document, the extent to which he was competent, unbiased, and in a position to report accurately the events described, where and when the document was written, and what other investigators may have discovered concerning its dependability. When comparable sources exist, they should be consulted in this connection. A source may be examined with reference to internal consistency. Contradictory or inconsistent statements within a source are indicative of a lack of dependability.

Records, catalogs, newspapers, diaries, personal letters, and other writings of the period may usually be accepted as dependable sources, but data taken from them should be correctly labeled. For example, if the factual statements relate to legislative enactments, they should not be used uncritically as descriptive of educational practice. Legislation pertaining to education sometimes legalizes practices long current. In other cases a law is enacted but observed to only a limited extent. This is illustrated by the Massachusetts Law of 1642. It directed officials of the towns to ascertain from time to time whether or not parents and masters of apprentices were fulfilling their educational duties. The Law of 1647 ordered the establishment of schools, indicating that the Law of 1642 was not functioning effectively. Hence, an inference from the Law of 1642 that parents and masters universally fulfilled their educational duties would be unjustified. Neither should one infer from the Law of 1647 that schools had not been established prior to 1647. Other sources indicate that a number of schools were in operation for years prior to that date.

Newspaper advertisements are an important source of data respecting the private schools of colonial days.¹ One can be

¹ See Seybolt, R. F. "The Evening School in Colonial America," *University of Illinois Bulletin*, Vol. 22, No. 31, *Bureau of Educational Research Bulletin*, No. 24. Urbana: University of Illinois, 1925. 68 pp.

Seybolt, R. F. "Source Studies in American Colonial Education. The Private School," *University of Illinois Bulletin*, Vol. 23, No. 4, *Bureau of Educational Research Bulletin*, No. 28. Urbana: University of Illinois, 1925. 109 pp.

certain from these advertisements what courses were *offered* by the private schoolmasters. One can *infer* the courses taught with reasonable justification, but not with certainty since the advertisements may not have stimulated enrollment in all of the courses listed. Numerous independent advertisements of a given subject, however, would appear to justify with practical certainty that the subject was taught at the times indicated by the dates of the newspapers. The schoolmasters would not have been likely to continue to advertise had there not been a demand for the courses offered.

The preceding paragraph illustrates an important principle in historical research. The occurrence of an event is established by the agreement of reports of independent and competent witnesses. The appearance of advertisements relating to a given course in newspapers of different towns at approximately the same time would seem to establish that the course was taught at that time.

Although official records are generally considered dependable, the critical worker will be alert to detect evidence of inconsistency or error. This is especially desirable in the case of school records. Our present methods of child and financial accounting are a recent achievement, and it has been found that comparatively recent school records were, in some cases, incomplete or misleading. Reminiscences and accounts of events within the lifetime of the author, but written sometime after their occurrence, are usually less dependable than diaries or personal letters. The author's memory is not likely to be complete and accurate, at least in regard to dates and details of events. He is also likely to interpret happenings of the past in the light of subsequent events. Official reports may involve errors. Occasionally the error may be due to intentional falsification, but more frequently the cause is inadequate information for the period covered by the report. Information concerning educational practices and conditions within a state or larger area is usually gathered by correspondence and it is difficult, even today, to secure a response of 100 per cent to an

official request. Furthermore, sometimes information furnished the compiler is not accurate or complete.

When the data are generalizations found in writings of the period, it must be remembered that until comparatively recently society was highly provincial and consequently that only a very few writers possessed accurate knowledge concerning educational conditions over a large area. Even when the generalization is restricted to a limited area, it was not usually based upon systematically collected data. Furthermore, it is important to know whether the writer was influenced by any bias or prejudice.

The dependability of historical research. By being systematic and by checking the copied items, a historical investigator can insure the accuracy of the items as judged by the sources from which they were obtained, but evaluation of their historical accuracy and validity is likely to involve judgment and hence this phase must be designated as subjective. The selection of sources and of the items copied from them is also subjective. Inferences made with respect to cause and effect relationships and generalizations respecting the prevalence of a practice or an idea cannot be other than subjective. The historical research worker cannot keep himself out of his research, if we interpret such research to include more than the mere collection of data. The point to be noted, however, is that the historical investigator should seek to know his data and to make interpretation in accord with their limitations. In other words, he should endeavor to maintain a scientific attitude.

Generalizations should be made with caution. A particularly dangerous type of generalization is that a given event was the first of its kind. For example, first use of the questionnaire as a means of collecting data has been credited to Sir Francis Galton in about 1875.¹ There are numerous evidences, however, that this instrument was used prior to Galton's

¹ Henderson, E. N. "Francis Galton," *Cyclopedia of Education*, Vol. 3. New York: The Macmillan Company, 1912, p. 4.

time.¹ The cautious investigator will report that the event is the earliest of which he has record. In generalizing with respect to the prevalence of a given practice at a given time it is desirable to support the generalization by citations to numerous and well distributed occurrences of the practice. Generalization in historical research requires, as in other types of research, all of the data or a representative sample of them. Some very misleading statements appear in current histories of education because of the hasty generalizations made by their authors on the basis of scanty and unrepresentative data.

A BIBLIOGRAPHY OF TYPICAL HISTORICAL EDUCATIONAL RESEARCH STUDIES

The student interested in the history of education or in conducting a historical research will find the following references helpful. A critical reading of several of these studies affords an excellent learning activity for engendering a knowledge of historical research techniques. Other historical studies can be located by consulting Monroe, Walter S., and Shores, Louis, *Bibliography and Summaries in Education* under "History of Education."

ABELSON, PAUL. "The Seven Liberal Arts," *Teachers College, Columbia University Contributions to Education*, No. 11. New York: Bureau of Publications, Teachers College, Columbia University, 1906. 150 pp.

BROWN, S. W. "The Secularization of American Education as Shown by State Legislation, State Constitutional Provisions, and State Supreme Court Decisions," *Teachers College, Columbia University Contributions to Education*, No. 49. New York: Bureau of Publications, Teachers College, Columbia University, 1912. 160 pp.

COLE, P. R. "Later Roman Education in Ausonius, Capella, and the Theodosian Code," *Teachers College, Columbia University Contributions to Education*, No. 27. New York: Bureau of Publications, Teachers College, Columbia University, 1909. 39 pp.

DEARBORN, N. H. "The Oswego Movement in American Education," *Teachers College, Columbia University Contributions to Education*, No. 183. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 191 pp.

¹ See Monroe, W. S., et al. "Ten Years of Educational Research, 1918-1927," *University of Illinois Bulletin*, Vol. 25, No. 51, *Bureau of Educational Research Bulletin*, No. 42. Urbana: University of Illinois, 1928, pp. 36-38.

- FITZPATRICK, E. A. "The Educational Views and Influence of DeWitt Clinton," *Teachers College, Columbia University Contributions to Education*, No. 44. New York: Bureau of Publications, Teachers College, Columbia University, 1911. 157 pp.
- GOOD, H. G. "The Sources of Spencer's 'Education,'" *Journal of Educational Research*, 13: 325-35, May, 1926.
- GWYNN, AUBREY. *Roman Education from Cicero to Quintilian*. Oxford: Clarendon Press, 1926. 260 pp.
- HANSEN, A. O. *Liberalism and American Education in the Eighteenth Century*. New York: The Macmillan Company, 1926. 317 pp.
- JACKSON, G. L. "The Development of School Support in Colonial Massachusetts," *Teachers College, Columbia University Contributions to Education*, No. 25. New York: Bureau of Publications, Teachers College, Columbia University, 1909. 95 pp.
- KEMP, W. W. "The Support of Schools in Colonial New York by the Society for the Propagation of the Gospel in Foreign Parts," *Teachers College, Columbia University Contributions to Education*, No. 56. New York: Bureau of Publications, Teachers College, Columbia University, 1913. 279 pp.
- KNIGHT, E. W. "The Influence of Reconstruction on Education in the South," *Teachers College, Columbia University Contributions to Education*, No. 60. New York: Bureau of Publications, Teachers College, Columbia University, 1913. 100 pp.
- MADDOX, W. A. "The Free School Idea in Virginia before the Civil War," *Teachers College, Columbia University Contributions to Education*, No. 93. New York: Bureau of Publications, Teachers College, Columbia University, 1918. 225 pp.
- NOBLE, S. G. "Early School Superintendents in New Orleans," *Journal of Educational Research*, 24: 274-79, November, 1931.
- RABENORT, W. L. "Spinoza as Educator," *Teachers College, Columbia University Contributions to Education*, No. 38. New York: Bureau of Publications, Teachers College, Columbia University, 1911. 87 pp.
- REIGART, J. F. "The Lancastrian System of Instruction in the Schools of New York City," *Teachers College, Columbia University Contributions to Education*, No. 81. New York: Bureau of Publications, Teachers College, Columbia University, 1916. 105 pp.
- ROBBINS, C. L. "Teachers in Germany in the Sixteenth Century. Conditions in Protestant Elementary and Secondary Schools," *Teachers*

College, Columbia University Contributions to Education, No. 52. New York: Bureau of Education, Teachers College, Columbia University, 1912. 126 pp.

TAYLOR, H. C. "Educational Significance of the Early Federal Land Ordinances," *Teachers College, Columbia University Contributions to Education*, No. 118. New York: Bureau of Publications, Teachers College, Columbia University, 1922. 138 pp.

TOTAH, K. A. "The Contribution of the Arabs to Education," *Teachers College, Columbia University Contributions to Education*, No. 231. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 105 pp.

UPDEGRAFF, HARLAN. "The Origin of the Moving School in Massachusetts," *Teachers College, Columbia University, Contributions to Education*, No. 17. New York: Bureau of Publications, Teachers College, Columbia University, 1908. 186 pp.

WELLS, G. F. "Parish Education in Colonial Virginia," *Teachers College, Columbia University Contributions to Education*, No. 138. New York: Bureau of Publications, Teachers College, Columbia University, 1923. 95 pp.

WOODY, THOMAS. "Early Quaker Education in Colonial Pennsylvania," *Teachers College, Columbia University Contributions to Education*, No. 105. New York: Bureau of Publications, Teachers College, Columbia University, 1920. 287 pp.

CHAPTER VII

CONSTRUCTING MEASURING INSTRUMENTS

A. GENERAL PRINCIPLES

The scope of measurement in education. When measurement in the field of education is mentioned, we commonly think merely of the administration of a test designed to measure either general intelligence or some segment of school achievement. This is an unfortunate restriction. For years a fundamental *credo* of research workers has been: Whatever exists at all exists in some amount and can be measured.¹ Although this ideal has not been fully realized, many ingenious measuring instruments have been devised and today we are able to measure, at least crudely, many things other than general intelligence and school achievement. These include personality traits, interests, attitudes, socio-economic status of homes, school buildings, quality of supervision, educational need, and the like.

A basic assumption.² It may seem platitudinous to point

¹ This statement is frequently attributed to Thorndike, but the present writers have been unable to locate it in his writings. The following quotations, however, express the essential ideas. "Whatever exists at all exists in some amount." "We have faith in whatever people now measure crudely by mere descriptive words, helped out by the comparative and superlative forms, can be measured more precisely and conveniently if ingenuity and labor are set at the task."

Thorndike, E. L. "The Nature, Purposes, and General Methods of Measurements of Educational Products," *The Seventh Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1918, p. 16.

² No attempt is made to consider the assumptions involved in the measurement of human abilities and traits. The following references represent two attempts to formulate assumptions in certain fields of measurement.

Tyler, Ralph W. "Assumptions Involved in Achievement-Test Construction," *Educational Research Bulletin*, 12: 29-36, February 8, 1933.

Hendrickson, Gordon. "Assumptions Underlying Personality Measurement," *Journal of Experimental Education*, 2: 243-49, March, 1934.

The following reference may also be read with profit in this connection:

out that in attempting the measurement of human abilities and traits, their existence is assumed, but this assumption means more than is commonly realized. Existence implies stability. Human performance varies ¹ even when conditions are considered to be unchanged and forgetting is a characteristic phenomenon. Hence, it may be inferred that human abilities and traits are not perfectly stable. To the extent that they fluctuate, the basic assumption of stability is not satisfied and their measurement is subject to a significant limitation.

The meaning of measurement as applied to human abilities and traits. In measuring distance, weight, and the like, the procedure may be described as "direct," meaning that the measurer deals "directly" with the magnitude whose measurement is desired. The most direct measurement of human abilities or traits is accomplished by securing the performance that the ability or trait makes possible or insures and then describing this performance in quantitative terms. In this sense the measurement of calculation skill in arithmetic or of ability in handwriting may be described as *direct*. Measurement, however, may be *indirect*. Temperature is measured by measuring the height of a column of liquid. In this case, indirect measures are valid because the correlation between temperature and the height of the column of liquid is approximately 1.00. The weight of children may be measured indirectly by measuring their height but obviously the measures obtained in this way will not be highly valid. Many of our measures of human abilities and traits are indirect. For example, when measuring the silent reading ability of sixth-grade children, we usually desire an index of their performance when they read assignments in history, literary selections, and the like. In a particular case, we may desire an index of their performance when they read a particular type of material in response to certain directions and under certain conditions such as reading an article in an encyclopedia

Lindquist, E. F., and Anderson, H. R. "'Achievement' Tests in the Social Studies," *The Educational Record*, 14: 198-256, April, 1933.

¹ See page 131.

to secure information relative to certain questions and doing this in a library where there are distracting influences. In such cases, scores on a typical silent reading test are indirect measures of the designated ability. In many cases the difference between the obtained performance and the one we are interested in is even more obvious. We use the performances given in response to true-false exercises or some other type of objective test as a means of securing measures of the ability to do something very different. Such indirect measures may be useful but frequently they involve relatively large variable errors of validity.

A test score should be thought of as a measure of ability to do, capacity to do, or tendency to do under certain conditions and at a certain time. Achievement is commonly thought of as having some degree of permanency. Usually the ability whose measurement is desired is that which will function at some date after the administration of the test.¹ Hence, the complete label of test scores as measures would include specifications in regard to the conditions under which the ability to do, capacity to do, or tendency to do is to function and the date on which the functioning is to occur. For example, spelling achievement might be defined as the ability to spell under the conditions of actual writing a week after the testing, assuming normal forgetting and only incidental learning. Hence, if the meaning implicitly associated with the scores yielded by a test were explicitly stated, the attached label would be of the following type: "numerical index of ability to under conditions at date, the intervening learning and forgetting being"

A measure of an ability or trait is a numerical index of a designated performance to be given under prescribed conditions, and hence the only essential requirement is that the ob-

¹ For two studies of retention at the high school level, see

Layton, E. T. "The Persistence of Learning in Elementary Algebra," *Journal of Educational Psychology*, 46-55, January, 1932.

Kennedy, L. R. "The Retention of Certain Latin Syntactical Principles by First and Second Year Latin Students after Various Time Intervals," *Journal of Educational Psychology*, 23: 132-46, February, 1932.

tained measures correctly discriminate between individuals whose future performances differ even when these differences are small. In other words, the scores made by pupils are to be considered satisfactory measures if the difference between them corresponds to the difference between the future performances of these pupils under the prescribed conditions.

Direct measurement is not essential but it should be noted that indirect measurement depends upon the stability of the relationship between what the test measures directly and the ability or trait whose measurement is desired. This relationship is not necessarily a fixed one. It depends upon the objectives towards which the instruction has been directed and it is likely to be changed by the repeated administration of a test of a given type. Hence, indirect measurement is based upon a relationship that is subject to change. This means that the validity of a given instrument is not necessarily a stable characteristic. In other words, a test that is highly valid in one situation is not necessarily equally valid in another. This point is especially important in connection with true-false, multiple-choice, and other types of tests constructed by teachers. Such instruments may yield highly valid indirect measures of achievement when first administered to a group of students but the validity is likely to decrease as the students direct their efforts to becoming able to respond to them unless the ability to respond to such exercises is recognized as a desirable educational objective.¹

General types of instruments for measuring human abilities and traits. The most common types of instrument for measuring human abilities and traits is the *performance test*, which requires the person, whose capacity or achievement is being measured,

¹ The widespread use of objective tests by teachers is undoubtedly contributing much to setting the objectives towards which students direct their efforts. This condition is probably making the tests increasingly less valid as measures of the desired achievement. For an elaboration of this point and comments upon certain undesirable results of the use of the objective tests, see Douglass, H. R. "The Effects of State and National Testing on the Secondary School," *School Review*, 42: 497-509, September, 1934.

For other points of view see Barr, A. S., and others. "A Symposium on the Effects of Measurement on Instruction," *Journal of Educational Research*, 28: 481-527, March, 1935.

to give a performance, usually a written one, which represents the functioning of that capacity or achievement or which is considered to be indicative of this capacity or achievement. *Psychological questionnaires*, which are used to obtain measures of general patterns of conduct, represent a second type of measuring instrument. The performance required consists of answers to questions concerning practices, beliefs, preferences, judgments, and the like. A third type of measurement procedure is *rating or estimating* with or without a formal scale on the basis of general acquaintance or systematic observation. To these three types of instruments there may be added observation (noting the frequency of performance of certain acts), interviewing, and laboratory and clinical techniques.¹

Basic problems in the construction of measuring instruments. Before the construction of a test is begun, it is necessary to determine what is to be measured. Ability in spelling, ability in silent reading, ability in arithmetic, and the like are commonly used as if such phrases designated defined achievements. Usually they do not. In fact, we know relatively little about the nature of the abilities we claim to measure. Within a given subject-matter field, achievement usually represents a combination of two or more types of controls of conduct. There is also considerable overlapping between achievements in different fields, due in part to the presence of general intelligence as a common factor. Beyond such general items, our information is limited.

The analysis of human abilities and traits and the identification of their elements are *basic problems* and logically their solution should precede the construction of measuring instruments. These problems are engaging the attention of a number of workers,² but it will be some time before we have the basic in-

¹ For a more comprehensive treatment of measurement procedures, see Symonds, P. M. *Diagnosing Personality and Conduct*. New York: Century Company, 1931. Chapters II, III, XI, XII, and XIV form an excellent reference supplementary to the present discussion. Symonds also deals with free association tests.

² See references to factor analysis in Chapter XI.

formation that we need in test construction. Meanwhile, a test-maker should attempt to specify as definitely as possible the nature of the ability or trait he desires to measure. It is especially important that he recognize and distinguish between the different types of achievement. In the measurement volume of the report of the Commission on the Social Studies, seven rubrics are indicated for this field: ¹ (1) exact information (memorized facts), (2) technical vocabulary, (3) ability to apply ideas and information to new situations, (4) skills, (5) interests, (6) attitudes, (7) ability to express. The learnings included vary with the subject-matter field but an analysis of this type indicates the general character of the achievement. A testmaker may not wish to measure all phases but he should explicitly recognize the restriction.

The general problems involved in constructing a performance test. Given the specifications of the abilities to be measured, the next problem is to determine the type of test to be constructed. With reference to the difficulty of the exercises of a test, their arrangement may be irregular, uniform, or scaled. In the first type, the arrangement of the exercises is not related to their relative difficulty. Tests of this type are illustrated by those prepared informally by teachers. In the second type the exercises are of equal difficulty, or approximately so. In the last type the exercises vary from very easy to very difficult and are arranged in ascending order of difficulty. Frequently, in such tests the increase in difficulty from exercise to exercise is approximately uniform. Variations of these types include spiral and cycle tests. In the former, different sub-tests may be arranged in order of increasing difficulty, the exercises within each sub-test being uniform in difficulty. In cycle tests, different types of exercises recur at regular intervals. In determining the type of test to be constructed, the testmaker should be guided by the nature of the measurement desired. If it is desired to secure a measure of the rate of the functioning of a group of

¹ Kelley, T. L., and Krey, A. C. *Tests and Measurements in the Social Sciences*. New York: Charles Scribner's Sons, 1934, p. 105.

skills, the test should be uniform in difficulty. When "power" is to be measured, a scaled test is generally employed.¹

The second problem is that of devising and selecting test exercises that will secure pupil performances satisfactory for the measurement of the specified ability or trait. It is necessary that the performance given in response to an exercise be observable. A second requirement is that the exercises be valid. It is desirable that the responses to the exercises be objectively scorable.

The third general problem relates to the determination of the length of the test (number of test items), organization of the test items, administration time, and formulation of an explanation of the nature of the exercises to the subjects taking test and of instructions for administering it.

The fourth problem relates to the quantitative description of the obtained test performance. Several questions are involved—rules for scoring, weighting of the exercises, units of the scale of description, and its zero point. In the case of handwriting, written composition, and other performances for which a general quality description is desired, there are the problems of devising an appropriate quality scale and determining the best technique for using it.

Testing the completed instrument forms the final problem. Two questions require consideration: (1) How accurately does the test measure what it actually measures? (2) How accurately does it measure the specified ability or trait? These two questions are commonly designated as those of *reliability* and *validity*.

In view of the large number of tests and other measuring instruments that have been constructed, it may appear that these problems should have been solved for many abilities and traits. It is true that for measuring general intelligence, general educational status below the senior high school level, and achievement in certain subject-matter fields, a number of tests

¹ For a more extended discussion, see Monroe, W. S. *An Introduction to the Theory of Educational Measurements*. Boston: Houghton Mifflin Company, 1923, pp. 62-76.

are available whose scores are highly reliable and exhibit fairly close agreement with the criterion measures used to determine their validity. Critical study of available tests, however, reveals a number of unsolved problems and the research worker who has occasion to select a measuring instrument for a particular purpose, especially when conducting an experimental investigation, seldom is able to find one that exactly meets his requirements. Hence, the field of measurement still offers challenges to the research worker. In fact, there are a number of important problems that have scarcely been touched.

B. DETAILS OF TEST CONSTRUCTION

Devising and selecting test exercises. The criteria to be observed in devising and selecting test exercises are (1) observability of performance, (2) convenience in administering the test and in scoring the test papers, (3) objectivity in scoring, (4) difficulty of the test items, and (5) validity. The second and third criteria are important requirements when the test is designed for general use, but when it is being devised as a means of securing measures needed in an experimental study or some other investigation, convenience in administering and in scoring the test papers and objectivity¹ are desirable but not necessary. Having determined upon the type or types of exercises to be employed, the testmaker proceeds to devise a number of such exercises being guided by his estimates of validity and difficulty. These exercises are then administered to a typical population for an experimental determination of validity and difficulty.

Criteria of validity to be observed in devising test exercises.² Although indirect measurement is possible and the only de-

¹ Many persons appear to consider objectivity in scoring an essential requirement for even a reasonably satisfactory test. This position is not justified by the facts. For example, see Traxler, A. E., and Anderson, H. A. "The Reliability of an Essay Test in English," *School Review*, 43: 534-39, September, 1935.

English essay examinations scored under carefully controlled conditions yielded reader reliabilities of .94 for one form and .85 for the other. Correlation between two forms was .60.

² The reader will find it helpful to refer to the discussion of the causes of errors of validity in Chapter V.

pendable determination of validity is by ascertaining the degree to which the exercise contributes to the identification of individual differences in the ability or trait whose measurement is desired, the testmaker should give attention to the content of the exercises. It is desirable to make the measurement as direct as possible. If specific habits are to be measured, it is sometimes possible to devise exercises that will call for the normal functioning of them. In the case of the calculation skills of arithmetic, there is the obvious suggestion that the test exercises consist of typical examples. In many cases, however, normal functioning is not feasible. When spelling ability is to be measured, the dictation of word lists does not provide for normal functioning. The act of silent reading does not eventuate in an observable performance. Hence, it is necessary to devise a type of exercise that will call for silent reading plus an observable performance indicative of the ability to read. When an exercise calling for normal functioning is not feasible, the testmaker should be guided by the experimental evidence pertaining to the relative validity of the possible types of exercises.

When the achievement to be measured consists of the ability to respond to arithmetical problems or to other types of thought questions, an obvious suggestion is that the test exercises should present typical problematic situations in the field of knowledge achievement being considered. In complying with this suggestion, it is necessary to bear in mind the requirement of convenience and objectivity in scoring. If the test is designed as an instrument for general use, there is the further requirement that the time necessary for responding to an exercise be reasonably short. In order to satisfy these requirements, testmakers have proposed various types of objective exercises—true-false, multiple-choice, completion, and the like. The use of such exercises raises the question of the extent to which they measure the achievement that functions in responding to questions that ask the student to discuss, explain, compare, and the like. They cannot measure such achievement directly, but indirect measurement is possible. On the basis of rather crude data, some test-

makers have concluded that tests consisting of objective exercises do measure such achievement to a sufficient extent to justify their use. More carefully planned inquiries indicate that an objective test measures about 60 per cent of what is measured by an essay type of examination when the latter is carefully prepared and scored.¹ A community of function of this degree does not make the measurement of knowledge achievement by means of objective tests very satisfactory. When a short testing time and convenience and objectivity in scoring are not imperative requirements, the use of exercises that will yield more direct measures of knowledge achievement is to be recommended.

When the achievement to be measured consists of general patterns of conduct, the nature of this rubric of controls of conduct suggests that the criterion of the acquisition of an attitude, interest, ideal, or other generalized control is the conformity to the pattern in responding to a variety of situations. This seems to be implied in the concept of a generalized control of conduct. There is the further implication of conformity to the pattern in spite of temptation to do otherwise, or at least of conformity in situations in which there is no expressed or implied requirement of conformity. Hence, a formal test administered within a typical time limit does not provide for the normal functioning of a general pattern of conduct. This means that the construction of a formal test to measure an attitude, interest, ideal, or other generalized control must be based upon the principle of indirect measurement, at least in part.

Criteria relating to the difficulty of test exercises. In the case of speed tests, especially when the function is narrow, the nature of the exercises is determined by the purpose of the measuring instrument. This would be true of a test designed to measure

¹ Cochran, R. E., and Weidemann, C. C. "'Explain' Essay versus Word-Answer Fact Test," *The Phi Delta Kappan*, 17: 59-61, December, 1934.

This point is considered further in Chapter VIII. The technique employed in these recent investigations is described in Chapter XI as well as the meaning of the statement that one test measures 60 per cent of what is measured by another test.

skill in the fundamental addition combinations or skill in doing arithmetical examples of a given type. In many cases, however, the difficulty of the exercises requires consideration.

Although a testmaker's estimates of difficulty will not be very dependable, he should endeavor to devise exercises whose difficulty is that desired in the completed test. In a scaled test, the exercises vary in difficulty from very easy to very difficult. Hence, when this type of test is being constructed, the testmaker should endeavor to devise exercises that represent a wide range of difficulty. If a uniform test is being constructed, he should endeavor to have the exercises approximately equivalent in difficulty and the degree of estimated difficulty should be that for which the test will yield the most valid measures. It is obvious that an exercise so easy that all subjects give the correct response or so difficult that none respond correctly is worthless for testing purposes, because its discriminating value would be zero. This condition suggests the inference that the maximum discriminating value is attained when an exercise is done correctly by about fifty per cent of the subjects. This inference is supported by some experimental evidence, but it appears that the discriminative value varies only slightly between thirty and seventy per cent of correct responses.¹ Hence, when devising exercises for a uniform test, an effort should be made to have the difficulty fall within these limits.

Experimental determination of the relative validity of individual test exercises. A valid test reveals differences in ability that exist and one that fails to reveal a difference which is at all marked is distinctly lacking in this quality. This concept of validity may be applied also to individual test items. When so applied, it means that in general the responses to a single exercise will be different for groups of pupils known to differ with respect to average status of the ability whose measurement is being attempted. For example, if two groups of pupils differ

¹ Thurstone, T. G. "The Difficulty of a Test and Its Diagnostic Value," *Journal of Educational Psychology*, 23: 335-43, May, 1932.

Symonds, P. M. "Choice of Items for a Test on the Basis of Difficulty," *Journal of Educational Psychology*, 20: 481-93, October, 1929.

with respect to silent reading ability, a valid exercise will yield a higher per cent of correct responses by the group whose average status is the higher. If the per cent of correct responses is approximately the same for the two groups, the validity of the exercise approaches zero. Another way of thinking of the validity or discriminating power of a test item is in terms of the accuracy with which subjects responding to the exercise are correctly placed on the scale of ability by their responses. Perfect placement is attained when all subjects who fail the item occupy positions on the scale below any subject who gives the correct response.

The first of the above expositions of the meaning of validity as applied to individual test items¹ suggests a method of determining the relative validity of the exercises of the preliminary form of a test. The essential requirement is two or more groups of subjects differing in average status with respect to the specified ability. For precise determinations it is necessary to have a series of groups whose average standings are evenly spaced over the range of the ability within which measurement is desired. For example, suppose ability to spell in normal writing is specified. The first step would be to secure a number of groups of pupils differing in average status with respect to this ability and evenly spaced on the scale of the ability. Suppose that ten such groups have been identified. Then the validity of a single test item would be indicated by the corresponding series of per cents of correct responses. A high degree of discriminating power would be indicated by a series of per cents exhibiting fairly uniform positive increments.

A criterion is essential for the determination of groups that differ with respect to the specified ability. In the case of general intelligence and achievement abilities that obviously increase from year to year or from grade to grade, sequential age or grade

¹ Sometimes "validity of test items" is used in the sense of *internal consistency*. This interpretation leads to the use of the total score on the test being taken as the criterion. Although investigation of the internal consistency of a test may be desirable, the procedure should not be thought of as a means of determining the validity of test items.

groups of typical pupils have been used. Binet, who in collaboration with Simon in 1905 devised and published the first general intelligence test,¹ employed age groups and selected those exercises that were responded to correctly by a larger per cent of the pupils of the older groups. Otis² in 1923 employed one group of retarded pupils and one of accelerated pupils. Thurstone³ selected groups on the basis of scholarship records. In validating exercises for trade tests, groups classified on the basis of training and experience as novices, apprentices, journeymen, and experts have been used. Groups selected in such ways are not likely to be very evenly spaced with reference to average status and there will be considerable overlapping. Hence, the determination of the validity of the test items will not be entirely satisfactory. A more precise criterion is needed for selecting spaced groups. If a test is available which experience has proven to yield reasonably valid measures of the ability or trait, it may be used for selecting the members of the several groups. The Stanford Revision of the Binet test is frequently used for this purpose in constructing a group intelligence test. Usually, however, such an instrument is not available.

After the test items have been administered to a series of

¹ Binet, A., et Simon, T. "Méthodes Nouvelles pour le Diagnostic du Niveau Intellectuel des Anormaux," *L'Année Psychologique*, 11: 191-244, 1905.

Revisions were published by Binet and Simon in 1908 and by Binet alone in 1911. In 1908 Goddard made a translation of the scale devised by Binet and in 1911 he published a revision. Kuhlmann published a revision in 1912. In the same year the Stanford Revision first appeared. This scale prepared by Terman and others was made generally available in 1916. In 1917 Otis, working under Terman, devised what is generally considered the first group intelligence scale. His work was adopted by the committee of psychologists who prepared the well-known Army Alpha Scale-used in the testing of our military forces in 1917-18. After the war, several group intelligence tests were prepared for school use and intelligence testing in the public schools became widespread. For a brief history of the development of intelligence tests consult the following reference or one of several texts on intelligence tests.

Monroe, Walter S., et al. "Ten Years of Educational Research," *University of Illinois Bulletin*, Vol. 25, No. 51, *Bureau of Educational Research Bulletin*, No. 42. Urbana: University of Illinois, 1928, pp. 89-90, 94-95.

² Otis, A. S. "The Making of a Classification Test," *Contributions to Education*, Vol. I. Yonkers-on-Hudson, New York: World Book Company, 1924, pp. 149-59.

³ Thurstone, L. L. "Cycle-Omnibus Intelligence Test for College Students," *Journal of Educational Research*, 4: 265-78, November, 1921.

spaced groups, there remains the task of calculating an index of the discriminating power of each item. Vincent¹ has reported a study in which an index of discriminating power of test items was obtained by calculating the per cent of pupils making a wrong response to the exercise who had criterion scores equal to or greater than the median of the criterion scores of the pupils doing the exercise correctly. Other techniques have been described by Paterson,² Symonds,³ Wilson, Welsh, and Gulliksen,⁴ Lentz, Hirshstein, and Finch,⁵ and Zubin.⁶ Cook⁷ has reported a study in which he evaluated five methods applicable to test items whose scoring is dichotomous. He claims that the bi-serial r technique⁸ is probably more reliable than the other methods he investigated experimentally and should be used where the items approach uniformity in difficulty or represent

¹ Vincent, Leona. "A Study of Intelligence Test Elements," *Teachers College, Columbia University Contributions to Education*, No. 152. New York: Bureau of Publications, Teachers College, Columbia University, 1924, pp. 9 f.

² Paterson, D. G. *Preparation and Use of New Type Examinations*. Yonkers-Hudson, New York: World Book Company, 1925, p. 60.

³ Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 535.

⁴ Wilson, W. R., Welsh, G., and Gulliksen, H. "An Evaluation of Some Informal Questions," *Journal of Applied Psychology*, 8: 206-14, June, 1924.

⁵ Lentz, T. F., Hirshstein, Bertha, and Finch, J. H. "Evaluation of Methods of Evaluating Test Items," *Journal of Educational Psychology*, 23: 344-50, May, 1932.

⁶ Zubin, Joseph. "The Method of Internal Consistency for Selecting Test Items," *Journal of Educational Psychology*, 25: 345-56, May, 1934.

⁷ Cook, W. W. "The Measurement of General Spelling Ability Involving Controlled comparisons between Techniques," *University of Iowa Studies, Studies in Education*, Vol. 6, No. 6. Iowa City, Iowa: University of Iowa, 1932. 112 pp.

The "Index of Discrimination D" described by Cook was used by Lindquist and Anderson. See

Lindquist, E. F., and Anderson, H. R. "Objective Testing in World History," *Historical Outlook*, 21: 115-22, March, 1930.

See also Lindquist, E. F., and Cook, W. W. "Experimental Procedures in Test Evaluation," *Journal of Experimental Education*, 1: 163-85, March, 1933.

Richardson, M. W. "Notes on the Rationale of Item Analysis," *Psychometrika*, 1: 69-75, March, 1936.

Richardson, M. W. "The Relation between the Difficulty and the Differential Validity of a Test," *Psychometrika*, 1: 33-49, June, 1936.

⁸ Bi-serial r represents the correlation between success and non-success on a given exercise and the criterion and measures. Several formulae have been proposed. See McNamara, W. J., and Dunlap, J. W. "A Graphical Method for Computing the Standard Error of Bi-serial r ," *Journal of Experimental Education*, 2: 274-77, March, 1934. One formula is given in Chapter IX, page 236.

high levels of ability.¹ When the scoring of the exercises of a test is not dichotomous, i.e., when the responses may be classified into more than two categories, three methods have been proposed—correlation ratio or eta, McCall method, and Long method.²

Although the techniques proposed for determining the validity of test items have not been described, it has perhaps occurred to the reader that the amount of labor involved is so enormous that precise test construction is not possible except for one who has a large and trained clerical staff at his command. This is true, but it should be noted that the number of hours of labor may be greatly reduced when calculating and tabulating machines are available. Lindquist and Cook³ state that for bi-serial r they devised a procedure by means of which a single operator using Hollerith tabulating equipment can compute this index for from 50 to 75 items per hour, not including the time required to punch the cards.

The obtained index of the discriminating power of a test exercise is a function of the criterion and the experience of the subjects to which it was administered as well as of the content of the item. If the criterion measures are lacking validity, the obtained indices will be in error. If the exercise is unknown to the subjects and they merely guess in responding to it, the validity of the item will be zero. If the subjects have been taught the wrong response, a negative validity will be obtained.

¹ This conclusion is in agreement with the findings of Barthelmess, H. M. "The Validity of Intelligence Test Elements," *Teachers College, Columbia University Contributions to Education*, No. 505. New York: Bureau of Publications, Teachers College, Columbia University, 1931, pp. 11 f.

The bi-serial r technique has been employed by Brigham in his evaluation of items for the scholastic aptitude test of the College Entrance Examining Board. See Brigham, C. C. *A Study of Error*. New York: College Entrance Examining Board, 1932, 384 pp.

² For a description of these methods, see Barthelmess, H. M. *Op. cit.*, pp. 11 f. This reference also reports a comparative study of these methods together with other methods useful when the scoring is dichotomous. A later study has been reported by Long, J. A. "Improved Overlapping Methods for Determining Validities of Test Items," *Journal of Experimental Education*, 2: 264-68, March, 1934.

³ Lindquist and Cook, *op. cit.*, p. 182.

In general, the discriminating power of a test exercise at a particular grade level depends upon the curriculum and the quality of the instruction. Hence, experimental determinations of validity should not be thought of as characteristics of the test items alone. They are subject to change. Determinations for one school population may not be correct for another.

Experimental determination of the difficulty of test items. The relative difficulty of test items for a given population of subjects is indicated by the per cents of correct responses, but when it is considered desirable to have the items of a test evenly spaced with reference to difficulty, it is necessary to secure the translation of these per cents into measures of difficulty expressed in terms of an ability unit. The procedure commonly employed is based on the assumption that in a large unselected group the distribution of the ability for which a test is being constructed is approximately that of the normal probability curve. In other words, if the members of a large unselected group were classified on the basis of measures of the ability, the frequencies thus obtained would form a normal distribution. A point on the scale of this distribution designates a degree or level of ability and its location may be specified in terms of the per cent of the group to the left of the point. For example, if 35 per cent are to the left of the point, it is located a distance of $.385\sigma$ to the left of the mean of the distribution.¹

If a test exercise has been administered to a large unselected population, the subjects responding correctly are assumed to be the ones whose ability is greater than those who fail to give the correct response. Hence, the per cent of correct responses defines a point on the scale of ability. Tables have been constructed which give the points on the scale of ability corresponding to various per cents of correct responses. This scale is usually expressed from an arbitrary zero point instead of the mean.

¹ This value is obtained from a table giving the areas under the normal curve corresponding to deviations from the mean. If 35 per cent are to the left, there are 15 per cent between the point and the mean. The deviation corresponding to this area is $.385\sigma$.

See pages 79-80 for discussion of the normal probability curve.

McCall ¹ has proposed a point at 5.0σ to the left of the mean, but 2.5σ and 3.0σ have been used.

Stability of determinations of validity and difficulty of test items. On page 174, it was noted that changes in the curriculum or instruction may change the validity of an item. The difficulty value obtained for a test exercise depends upon the general status of the population to which it has been administered and upon the training the subjects have received relative to the exercise. Hence, determinations of the validity and difficulty of test items should not be thought of as stable characteristics. This principle means that a test item which is "good" for one population may be "poor" for another and that an item which is "good" at the time of the construction of a test may be "poor" at a later date. It is also possible that "poor" items may become "good" items. It is probably true that the quality of test items does not often change very rapidly, but determinations of validity and difficulty should not be thought of as highly stable. Furthermore, the validity index obtained for an item is probably influenced by the context in which it appears. Hence, the validity of an item in the final form of the test may not be the same as in the preliminary form.

Validity of items versus validity of test. The calculation of a validity index for a large number of items and the selection of those having the highest indexes for the test makes the process of test construction impressive, and many testmakers appear to have assumed that the procedure insures the best test that can be assembled from the items. The prime consideration in test construction is that the resulting instrument yield highly valid and reliable measures and it does not necessarily follow that the items with the highest validity indexes will make the most valid test. Investigation ² indicates that a test consisting of the "best"

¹ McCall, W. A. *How to Measure in Education*. New York: The Macmillan Company, 1922, pp. 274-75.

See also Monroe, W. S. *An Introduction to the Theory of Educational Measurements*. Boston: Houghton Mifflin Company, 1923, pp. 96-97.

² Smith, Max. "The Relationship between Item Validity and Test Validity," *Teachers College, Columbia University Contributions to Education*,

items may not be more valid than one consisting of "mediocre" items. These findings raise the question of the value of elaborate difficulty and validity analyses of test items. It is likely that selection of items on the basis of simple procedures will usually result in a test validity little lower than that attained by the elaborate procedures described.

C. DESCRIPTION OF TEST PERFORMANCES

Quantitative description of objectively scorable test performances. In formulating directions for scoring the responses to the exercises of a test and for combining the credits into a point score, attention should be given to (1) dimensions of ability, (2) rate units versus work units, (3) weighting, or relative credit for correct responses to the various exercises, (4) correction for guessing.

1. *Recognition of dimensions of ability.* A complete description of a test performance requires determination of its quality or accuracy, the difficulty of the exercises, and the rate at which the performance was given. In other words, a performance may be thought of as three dimensional. Little attention has been given to the dimensions of test performances, but precise measurement seems to require it. The number of exercises right is generally taken as the score. This score represents an unknown combination of two or all three of the dimensions. Under the caption of the Law of the Single Variable, Burgess suggested that the form of the test and its plan of administration be such that the difficulty and rate of work will be constant in all pupil performances.¹ This condition is approximated in a timed sentence dictation spelling test. The Courtis Silent Reading

No. 621. New York: Bureau of Publications, Teachers College, Columbia University, 1934. 40 pp.

¹ Burgess, May Ayres. *Measurement of Silent Reading Ability*. New York: Russell Sage Foundation, 1921, p. 61.

For one proposal for combining speed and quality in the case of handwriting, see

Gates, A. I. "The Relation of Quality and Speed of Performance: A Formula for Combining the Two in the Case of Handwriting," *Journal of Educational Psychology*, 15: 129-44, March, 1924.

Test provides for measuring rate and quality of comprehension independently. A few other testmakers have provided for separate scores to describe different aspects of the performance. It seems reasonable that recognition of the Law of the Single Variable might lead to significant refinements in measuring human abilities and traits. At least this detail of test construction offers a challenge to the critical student of educational measurement.

2. *Rate units versus work units.* In the case of a speed test, the performance may be described in terms of the number of units of work done (usually the number of exercises done correctly) within the allowed period of time, or in terms of the mean number of seconds per unit of work. The two methods of description yield different results and, hence, may lead to conflicting conclusions, particularly in research on the effect of practice on individual differences.¹ In such cases, it is advisable to express the measures in terms of both units of work and units of time.²

3. *Weighting or relative credit for correct responses to the various exercises.* If the exercises of the test are approximately equal in difficulty, the usual practice is to assign the value of one unit to each. When they vary materially in difficulty, a number of testmakers have devised a plan of weighting, usually based upon a

¹ Peterson, Joseph, and Barlow, M. C. "The Effects of Practice on Individual Differences," *The Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1928, pp. 211-30.

Peterson states on page 214 of the above reference: "The confusion resulting from comparing absolute gain in *amount of work done per unit of time* with *amount of time required to do a given piece of work*, I clearly pointed out and illustrated copiously with graphs ten years ago, but the warning then sounded and later emphasized has been heeded by only a few investigators." The writings referred to are:

Peterson, Joseph. "Experiments in Ball-Tossing: The Significance of Learning Curves," *Journal of Experimental Psychology*, 2: 178-224, 1917.

Peterson, Joseph. "Thurstone's Measure of Variability in Learning," *Psychological Bulletin*, 15: 452-55, 1918.

Peterson, Joseph. "Johnson's Measurement of Rate of Improvement under Practice," *Journal of Educational Psychology*, 15: 271-75, May, 1934.

² Conversion from units of work to units of time can easily be accomplished by computation of "harmonic means." See Odell, C. W. *Educational Statistics*. New York: The Century Company, 1925, pp. 97-104.

measure of the relative difficulty of the several exercises. In the case of tests of the scaled type, a pupil will not ordinarily do all of the exercises up to a given point of the scale and then fail in all beyond this point. The typical performance is one in which there is a scattering of correct responses beyond the point of the first incorrect one. This situation has caused some testmakers to devise a technique for computing a difficulty score which involves weighting.¹ It appears, however, that a system of weighting does not affect the reliability of test scores to a marked degree and the usual practice is to take the number of exercises done correctly as the pupil's score even though there may be marked variations in difficulty.²

Most of those who have studied the effect of weighting have employed reliability as the criterion. The effect of weighting upon the validity of the measures has received relatively little attention. Furthermore, it appears that the social value of an

¹ Kelley, T. L. "Thorndike's Reading Scale, Alpha 2, Adapted to Individual Testing," *Teachers College Record*, 18: 253-60, May, 1917.

Kelley, T. L. "A Simplified Method of Using Scaled Data for Purposes of Testing," *School and Society*, 4: 34-37, July, 1916.

Van Wagenen, M. J. "Table for Computing Mean Individual Scores in Educational Scales," *Teachers College Record*, 21: 441-51, November, 1920.

² The interested reader should consult the following reports of research on the problem. Corey contends that when scores are transmuted into grades that weighting has a significant influence on the grades assigned. Odell, Peatman, and Potthoff and Barnett contend that the effect of weighting is of some, but not of great significance in its effect on grades.

Corey, S. M. "The Effect of Weighting Exercises in a New Type of Examination," *Journal of Educational Psychology*, 21: 383-85, May, 1930.

Douglass, H. R., and Spencer, P. L. "Is It Necessary to Weight Exercises in Standard Tests?" *Journal of Educational Psychology*, 14: 109-12, February, 1923.

Monroe, W. S. "The Description of the Performances of Pupils on Exercises of Varying Difficulty," *School and Society*, 15: 341-43, March 25, 1922.

Odell, C. W. "Further Data Concerning the Effect of Weighting Exercises in New Type Examinations," *Journal of Educational Psychology*, 22: 700-04, December, 1931.

Peatman, J. G. "The Influence of Weighted True-False Test Scores on Grades," *Journal of Educational Psychology*, 21: 143-47, February, 1930.

Potthoff, E. F., and Barnett, N. E. "A Comparison of Marks Based upon Weighted and Unweighted Items in a New Type Examination," *Journal of Educational Psychology*, 23: 92-98, February, 1932.

Scates, D. E., and Noffsinger, F. R. "Factors Which Determine the Effectiveness of Weighting," *Journal of Educational Research*, 24: 280-85, November, 1931.

exercise should receive consideration as well as its difficulty. Frequently these criteria are incompatible as may be shown by considering two exercises, the first, calling for the date of the signing of the Declaration of Independence, and the second, calling for the date of a minor battle of the Civil War. The first exercise calls for information of much greater social value than the second. One would be considered ignorant indeed if he did not know the significance of the date, 1776. On the other hand, frequent contact with the date and its significance makes the first exercise a very easy one for most eighth-grade pupils. The second date is of much less social value, but it would probably be known by very few pupils. On the basis of the criterion of social value, the first exercise merits the greater weight, but on the basis of difficulty, the second exercise would be given the greater weight.

4. *Correction for guessing.* When the number of possible responses to an exercise is limited, it is possible to make the correct response merely by guessing. Hence, if the subjects to whom the test is administered guess when they do not know the response to make, the "number right" will tend to be too large as a score. The formula usually used to secure a corrected score is

$$\text{Score} = \text{Number right} - \frac{\text{Number wrong}}{(n - 1)}$$

The symbol, n , stands for the number of possible responses to an exercise. In the case of true-false and other alternative test exercises, the formula becomes:

$$\text{Score} = \text{Number right} - \text{Number wrong}.$$

For exercises requiring a choice from three possible answers, the formula becomes:

$$\text{Score} = \text{Number right} - \frac{\text{Number wrong}}{2}$$

When the number of possible responses is increased to five or more, the correction is probably not essential. After reviewing

the research relating to correction for guessing, Ruch ¹ concluded that the question of the effect upon reliability is still debatable, but that correction for guessing appears to increase the validity of the scores. It is likely that the effect of correction for guessing is conditioned by the character of the exercise and the instructions relative to guessing.²

Quantitative description of recorded performances by means of quality scales. Performances such as samples of handwriting or written compositions cannot be scored as "right" or "wrong." They must be described in terms of degrees of quality. This description may be accomplished by constructing a quality scale and then matching the performances with a step of this scale. The essential steps in the construction of a quality scale are as follows: (1) Collecting sample performances which represent as wide a range of quality as possible; (2) Selecting from this collection a limited number of samples (10 to 20) which represent the entire range of quality, and which differ from each other by approximately equal increments of quality; (3) Determining the quantitative description of the quality of each sample with reference to an established zero point, and arranging the samples in the form of a scale.

The procedures for the second and third of these steps have been described in several texts on educational measurement ³ and need not be considered here. The merit of the resulting quality scale depends upon the original collection of sample performances. Insofar as possible they should differ only with respect to the quality to be measured. For example, if the scale is to measure handwriting, the children should write the same text and employ the same style of writing. If the scale is to measure composition, the pupils should write on the same topic. Another consideration is that all degrees of excellence should be

¹ Ruch, G. M. *The Objective or New-Type Examination*. Chicago: Scott, Foresman and Company, 1929, pp. 318-57.

² See Wood, E. P. "Improving the Validity of Collegiate Achievement Tests," *Journal of Educational Psychology*, 18: 18-25, January, 1927.

³ For example, see Monroe, W. S. *An Introduction to the Theory of Educational Measurement*. Boston: Houghton Mifflin Company, 1923, pp. 135 f.

represented. For example, in the case of drawing, one sample should approximate zero drawing ability. It should be characterized as an attempt to draw, but an attempt which represents no accomplishment. The number of samples collected should be large so that one can be reasonably sure that all degrees of ability are represented.

The matching of the test performances with the steps of the scale is a subjective process which contributes to the unreliability of the measures secured. Several investigators have compared ratings by means of a scale with estimates of quality made without the use of a scale. The evidence is conflicting and hence disappointing. Odell, after summarizing investigations of the reliability of ratings made by means of composition scales and considering data with respect to the reliability of a set of scales for rating pupils' answers to thought questions, concluded that "if the scales themselves possess high merit and if those who employ them do so after the best possible preparation and in the best possible manner," the reliability of the resulting measures of quality will be higher than that of estimates made without the use of a scale.¹

Derived scales of measurement. The scale of measurement on which the point scores are expressed is determined by the structure of the test and the manner of computing the score. Usually the zero point is arbitrary and does not represent "not any of the thing measured." McCall has proposed a scale of measurement having the zero point at 5.0σ below the mean score of a large and unselected population of 12-year-old children. The unit of measurement is $.1\sigma$ and the measures expressed on this scale are known as T-scores.² Test scores can be converted

¹ Odell, C. W. "The Use of Scales for Rating Pupils' Answers to Thought Questions," *University of Illinois Bulletin*, Vol. 26, No. 36, *Bureau of Educational Bulletin*, No. 46. Urbana: University of Illinois, 1929, p. 28.

² For an explanation of the principle involved, see pages 82-83.

Ruch and Stoddard have pointed out that this principle was implied in Galton's treatment of correlation.

Ruch, G. M., and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers-on-Hudson, New York: World Book Company, 1927, p. 351.

For discussion of the T-score procedure, the interested reader should consult

into a number of other types of derived measures. Transmutation into age scores is a common procedure. Familiar examples are mental age, educational age, achievement age, arithmetic age, and reading age. A given reading age of 115 months means that the pupil who obtained this score did as well on the test as the median pupil whose chronological age is 115 months or 9 years 7 months.¹ When the median scores for a series of grades and for each month of the school year have been determined, point scores may be transmuted into grade scores. A grade score of 5.6 obtained by a pupil would indicate that his achievement on the test was equivalent to the median pupil in the sixth month of the fifth grade. The use of age scores and grade scores rests on the assumption that the groups whose scores are compared have been subjected to comparable educational experience. That is to say, the pupils have been subjected to the same curriculum and school organization.

the references given below. In addition to references dealing with McCall's procedure, some are included relative to the extension of McCall's procedure to several grade groups by Pintner and by Thurstone. Holzinger's criticisms and Thurstone's replies to his critic are included. The Thurstone and Holzinger references are grouped together in the order in which they appeared.

Kelley, T. L. "Comparable Measures," *Journal of Educational Psychology*, 5: 589-95, December, 1914.

McCall, W. A. "A Proposed Uniform Method of Scale Construction," *Teachers College Record*, 22: 31-51, January, 1921.

Monroe, W. S. *An Introduction to the Theory of Educational Measurements*. Boston: Houghton Mifflin Company, 1923, pp. 150-52.

Woodworth, R. S. "Combining the Results of Several Tests: A Study in Statistical Method," *Psychological Review*, 19: 97-123, March, 1912.

Thurstone, L. L. "A Method of Scaling Psychological and Educational Tests," *Journal of Educational Psychology*, 16: 433-51, October, 1925.

Thurstone, L. L. "The Scoring of Individual Performance," *Journal of Educational Psychology*, 17: 446-57, October, 1926.

Thurstone, L. L. "The Unit of Measurement in Educational Scales," *Journal of Educational Psychology*, 18: 505-24, November, 1927.

Holzinger, K. J. "Some Comments on Professor Thurstone's Method of Determining the Scale Values of Test Items," *Journal of Educational Psychology*, 19: 112-17, February, 1928.

Thurstone, L. L. "Comment by Professor L. L. Thurstone," *Journal of Educational Psychology*, 19: 117-24, February, 1928.

Holzinger, K. J. "Reply to Professor Thurstone," *Journal of Educational Psychology*, 19: 124-26, February, 1928.

Thurstone, L. L. "Scale Construction with Weighted Observations," *Journal of Educational Psychology*, 19: 441-53, October, 1928.

¹ For further discussion of age scores, see Monroe, *op. cit.*, pp. 155-56.

The computation of the percentile points of a distribution of scores makes possible the transmutation of scores into percentile measures. For example, if a score of 83 falls within the interval between the 47 and 48 percentile, it may be expressed as a percentile score of 47. The pupil's ability may be described as of the 47 percentile.¹ Although percentile scores are expressed in terms that can be easily understood, they do not possess the qualities which are possessed by derived scores expressed in terms of the variability of a chronological age group. The unit is not constant. The distribution is divided so that equal areas are obtained in calculating percentile scores. The divisions do not mark equal distances on the base line of the distribution. The result is that the difference between the degrees of ability represented by a 45 percentile score and a 50 percentile score is not at all equal to the difference between the degrees of ability represented by a 90 percentile score and a 95 percentile score. The latter is much larger. Percentile scores are useful, however, when precise comparison is not attempted, and at the high school level where age and grade scores are not applicable.

Composite score of a battery of tests. When the measuring instrument consists of several sub-tests and a single composite score is desired, it is necessary to derive a procedure for computing it. If it is desired to give equal weight to the measures resulting from the several sub-tests, the scores may be reduced to a common basis² and then added. If criterion measures are available, the weights for the maximum validity may be obtained by calculating the multiple regression equation.³

The psychological questionnaire.⁴ The psychological questionnaire is used to measure human traits (attitudes, interests, ideals, and other general patterns of conduct) by asking the

¹ For further discussion of percentile scores, see Monroe, *op. cit.*, p. 154.

Otis, A. S. *Statistical Method in Educational Measurement*. Yonkers-on-Hudson, New York: World Book Company, 1925, pp. 24-29, 95-100, 124-131. Ruch and Stoddard, *op. cit.*, pp. 347-50.

² See pages 82 f.

³ See pages 324 f.

⁴ For a more extended discussion see Symonds, P. M. *Diagnosing Personality and Conduct*. New York: The Century Company, 1931, pp. 122 f.

pupil questions relative to what he has done, how he is accustomed to do certain things, his likes, beliefs, choices, wishes, preferences, interests, and the like. The distinction between it and a performance test is that the latter attempts to secure the functioning of the ability whose measurement is desired or the functioning of a closely correlated ability. In the psychological questionnaire there is no such attempt. It merely asks questions whose answers experience has shown to be indicative of the status of the trait whose measurement is desired. Some of the questions call for deliberation and the expression of a judgment. Others ask for immediate expression of choices or other reactions. Usually the questions are expressed in a form such that the scoring of the answers is objective.

The only essential criterion for devising and selecting the items of a psychological questionnaire is their effectiveness in indicating individual differences in the trait whose measurement is desired. A question may appear to be inconsequential or even ridiculous,¹ but if experience shows that it is effective, its appropriateness has been demonstrated. In general, however, a superior initial selection of questions will be secured by analyzing the behavior associated with the trait and attempting to formulate pertinent questions.

A psychological questionnaire may be used as a means of securing a controlled interview for the purpose of diagnosing children individually. When used for this purpose by a competent person, a definite plan of scoring may not be advisable, but when groups of subjects are being studied, a numerical score is usually desired. The criterion of the validity of a method of scoring is the effectiveness of the resulting measures in revealing individual differences in the trait whose measurement is desired. In general no answer can be considered as wrong.

¹ For illustrations see Wells, F. L. "Report on a Questionnaire Study of Personality Traits with a College Graduate Group," *Mental Hygiene*, 9: 113-27, January, 1925.

Symonds, P. M. "A Studiousness Questionnaire," *Journal of Educational Psychology*, 19: 152-67, March, 1928.

The questions do not call for facts as in the case of performance tests. Any response to an item of a psychological questionnaire may be significant and in the present stage of the development of this field of measurement a person constructing a measuring instrument of this type should try out several methods of scoring, at least all promising ones, and calculate the correlation between the resulting sets of scores and the criterion.

A number of psychological questionnaires were used in the comprehensive investigation of Terman and his co-workers, "Mental and Physical Traits of a Thousand Gifted Children."¹ In Chapters XIII, XIV, and XV are described instruments for the measurement of scholastic, occupational, play, and reading interests. In Chapters XVII and XVIII are described questionnaires for the measurement of character and personality traits. Voelker's ten tests of trustworthiness, Cady's measures of incorrigibility, and the Woodworth-Cady Questionnaire for the measurement of emotional stability are also described. The Woodworth-Cady Questionnaire is reproduced in full on pages 501 to 505. An illustration of the efforts to measure the less tangible and definite controls of conduct is afforded by an attempt to measure "faith in God."²

¹ Terman, L. M., et al. "Mental and Physical Traits of a Thousand Gifted Children," *Genetic Studies of Genius*, Vol. I. Stanford, California: Stanford University Press, 1925. 648 pp.

² Donnelly, Harold I. "Measuring Certain Aspects of Faith in God as Found in Boys and Girls Fifteen, Sixteen, and Seventeen Years of Age," *A Thesis in Education Presented in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy*. Philadelphia: University of Pennsylvania, 1931. 118 pp.

The reader interested in psychological questionnaires may find the following references helpful:

Droba, D. D. "A Scale of Militarism—Pacifism," *Journal of Educational Psychology*, 22: 96–111, February, 1931.

Garretson, O. K. "Relationships between the Expressed Preferences and the Curricular Abilities of Ninth-Grade Boys," *Journal of Educational Research*, 23: 124–32, February, 1931.

Lincoln, E. A., and Shields, F. J. "An Age Scale for the Measurement of Moral Judgment," *Journal of Educational Research*, 23: 193–97, March, 1931.

Thurstone, L. L. "A Scale for Measuring Attitude toward the Movies," *Journal of Educational Research*, 22: 89–94, September, 1930.

Watson, Goodwin. "Happiness among Adult Students of Education," *Journal of Educational Psychology*, 21: 79–111, February, 1930. An example of measurement by self-estimate.

D. ESTIMATING THE EXCELLENCE OF MEASURING INSTRUMENTS

The reliability and validity of a measuring instrument. After a measuring instrument has been constructed including the formulation of directions for administering it and for obtaining the score, it remains to determine how accurately the instrument measures the ability or trait specified by its expressed or implied function. In Chapter V four types of errors were noted: (1) variable errors of measurement, (2) variable errors of validity, (3) systematic errors of measurement, (4) systematic errors of validity. Since the systematic error in the scores yielded by a test fluctuates with the manner of its administration and other conditions not connected with its structure, indices of only variable errors of measurement and variable errors of validity can be associated with a measuring instrument. The reliability of a test has reference to the variable errors of measurement and the validity of a test to the variable errors of validity.¹ When used in a technical sense, these terms have no connection with systematic errors.

The association of an index of variable errors of measurement or of the variable errors of validity with a test implies that the "average" magnitude of the errors to be expected in the case of a given test is a fixed characteristic of the measuring instrument. This implication is not necessarily true. In the case of a particular administration of a given test, the actual variable errors of measurement or the actual variable errors of validity may be materially greater than the index indicates. It is also possible that they may be less.² Hence, the indices dealt with in the following pages should be thought of as representing estimates of the magnitude of the variable errors to be expected rather than as fixed characteristics of a test or other measuring instrument.

¹ Although "validity" may be defined so that it refers to only variable errors of validity, a calculated index of validity usually is a measure of variable errors of measurement and variable errors of validity.

² For a study of the effect of practice upon the reliability of a test, see Anastasi, Anne. "The Influence of Practice upon Test Reliability," *Journal of Educational Psychology*, 25: 321-35, May, 1934.

Determining an index of the magnitude of the variable errors of measurement to be expected in the scores yielded by a given instrument. The determination of an index of the magnitude of the variable errors of measurement ¹ to be expected in the measures obtained by means of a given instrument is based upon two independent sets of scores for a group of subjects representative of the population for which the test is designed. Three methods are used for securing these scores: (1) repeating the administration of the test, (2) administering two forms of the test, and (3) securing two series of scores by dividing the test into two equivalent parts, usually by using the odd numbered items as the basis of one score and the even numbered ones as the basis of the other. None of these methods is entirely satisfactory. Both sets of scores should be typical of testing conditions that may be expected to prevail when the test is administered to a group of subjects. When a test is readministered, the conditions are likely not to be typical because the subjects may remember some of the exercises and their responses. Their attitude may also be affected. When two forms of a test are used, there is likely to be some "transfer" from the first to the second testing. Furthermore, unless very carefully constructed, the two forms are likely not to be entirely equivalent. When the two sets of scores are obtained by dividing a test into halves, the testing conditions are identical and hence the effect of typical variations in testing conditions is eliminated. Furthermore, this procedure introduces other difficulties that will be noted later.

1. *Coefficients of reliability.* The term "reliability coefficient" appears to have been used first by Spearman in 1910. Reliability coefficients, however, appeared in his writings as early as 1904 and were defined as "the average correlation between

¹ The variable error of measurement is the difference between an obtained score and the corresponding theoretical true score (X_{∞}) which is defined as the mean of a large number of scores made by the same subject on equivalent forms of the test, each score being corrected for systematic error. See pages 130-32 for discussion of causes of variable errors of measurement.

one and another of these several independently obtained series of values." ¹ In other words, the basis of a reliability coefficient was to be two independently obtained sets of measures of the same thing. This is equivalent to saying that the basis should be two sets of measures the pairs of which differ only because of the presence of variable errors of measurement. A reliability coefficient does not measure any systematic error that may be present.² When a coefficient of reliability is calculated from IQ's or other quotient scores, "spurious correlation" is introduced. For example, if an intelligence test with zero reliability is administered to a population heterogeneous with reference to chronological age, the reliability of the IQ's will be .50. For references dealing with "spurious correlation" see Chapter XI, page 388.

The coefficient of correlation between the two sets of paired scores obtained by either of the first two of the above procedures is commonly called the coefficient of reliability (r_{1r}).³ As pointed out above, the paired measures obtained either by repeating a test or by using two forms of a test do not completely qualify as "two independently obtained sets of measures of the same thing." In general, the coefficient calculated from scores obtained by repeating the test is likely to be slightly too large and one calculated from scores obtained from duplicate forms slightly too small.

¹ Spearman, C. "The Proof and Measurement of Association between Two Things," *American Journal of Psychology*, 15: 90-101, January, 1904.

Spearman, C. "General Intelligence Objectively Determined and Measured," *American Journal of Psychology*, 15: 253-93, April, 1904.

For a definition of reliability in terms of variance ratio, see Dunlap, J. W. "Comparable Tests and Reliability," *Journal of Educational Psychology*, 24: 442-53, September, 1933. This is an excellent critical discussion of the techniques for determining reliability. For an explanation of the variance ratio see Chapter XI.

² For an illuminating reference on this point, see Daniel, R. P. "Basic Considerations for Valid Interpretations of Experimental Studies Pertaining to Racial Differences," *Journal of Educational Psychology*, 23: 15-27, January, 1932.

³ Kelley favors the use of the term "retesting coefficient" for a coefficient obtained by the first procedure; see Kelley, T. L. *Interpretation of Educational Measurements*. Yonkers-on-Hudson, New York: World Book Company, 1927, pp. 39-40. "Consistency coefficient" has also been proposed.

The Spearman-Brown formula ¹ is employed as a means of calculating a coefficient of reliability from the coefficient of correlation between the two sets of scores obtained by the third procedure. Using the symbol, $r_{\frac{1}{2} I, \frac{1}{2} II}$, to designate the coefficient of correlation between the two halves of the test, the formula is

$$r_{II} = \frac{2r_{\frac{1}{2} I, \frac{1}{2} II}}{1 + r_{\frac{1}{2} I, \frac{1}{2} II}}$$

The use of this formula has been questioned. Shen ² has

¹ The formula was published simultaneously by Spearman and by Brown:

Spearman, C. "Correlation Calculated from Faulty Data," *British Journal of Psychology*, 3: 71-295, 1910.

Brown, W. "Some Experimental Results in the Correlation of Mental Abilities," *British Journal of Psychology*, 3: 296-322, 1910.

² Shen, Eugene. "The Standard Error of Certain Estimated Coefficients of Correlation," *Journal of Educational Psychology*, 15: 462-65, October, 1924. See also the criticism by Holzinger and the reply of Shen:

Holzinger, K. J., and Clayton, Blythe. "Further Experiments in the Application of Spearman's Prophecy Formula," *Journal of Educational Psychology*, 16: 289-99, May, 1925.

Shen, Eugene. "A Note on the Standard Error of the Spearman-Brown Formula," *Journal of Educational Psychology*, 17: 93-94, February, 1926.

See also

Lanier, L. H. "Prediction of the Reliability of Mental Tests of Special Abilities," *Journal of Experimental Psychology*, 18: 69-113, February, 1927.

Holzinger, K. J. "Note on the Use of the Spearman-Brown Prophecy Formula for Reliability," *Journal of Educational Psychology*, 14: 302-05, May, 1923.

Douglass, H. R., and Cozens, F. W. "On Formula for Estimating the Reliability of Test Batteries," *Journal of Educational Psychology*, 20: 369-77, May, 1929.

Farnsworth, P. R. "The Spearman-Brown Prophecy Formula and the Seashore Tests," *Journal of Educational Psychology*, 19: 586-88, November, 1928.

Kelley, T. L. "The Applicability of the Spearman-Brown Formula for the Measurement of Reliability," *Journal of Educational Psychology*, 16: 300-03, May, 1925.

Kelley, T. L. "Note on the Reliability of a Test: A Reply to Dr. Crum's Criticism," *Journal of Educational Psychology*, 15: 193-204, April, 1924.

Remmers, H. H. "The Equivalence of Judgments in the Sense of the Spearman-Brown Formula," *Journal of Educational Psychology*, 22: 66-71, January, 1931.

Remmers, H. H., Shock, N. W., and Kelly, E. L. "An Empirical Study of the Validity of the Spearman-Brown Formula as Applied to the Purdue Rating Scale," *Journal of Educational Psychology*, 18: 187-95, March, 1927.

Ruch, G. M., Ackerson, Luton, and Jackson, J. D. "An Empirical Study of the Spearman-Brown Formula as Applied to the Educational Test Material," *Journal of Educational Psychology*, 17: 309-13, May, 1926. (Continued next page)

shown that the standard error of the estimated coefficient is greater than that of an equivalent reliability coefficient calculated directly. In practice, however, this condition is likely to be less significant than failure to satisfy completely the assumptions on which the Spearman-Brown formula is based. The formula is a special case of the one for the correlation of sums ¹ which simplifies to the above form only when certain conditions are satisfied. The reliability coefficient r_{1I} may be written $r(\frac{1}{2} + \frac{1}{2})(\frac{I}{II} + \frac{I}{II})$. When expressed in this form, four sets of measures are apparent, those from the two halves of Test 1 and those from the two halves of Test I. Two assumptions are made: (1) the standard deviations of these four sets of measures are equal; (2) the various intercorrelations are equal. In applying the Spearman-Brown formula, one has only two sets of measures, one designated as $X_{\frac{1}{2}}$ and the other as $X_{\frac{I}{II}}$. If the standard deviations of these two sets of measures are equal, the r_{1I} obtained is the coefficient of correlation between two hypothetical tests, one formed by doubling the half designated as $\frac{1}{2}$ and the other formed by doubling the half designated as $\frac{I}{II}$. The obtained r_{1I} is the correct r_{1I} provided the measures resulting from the added portions have the same standard deviations as the measures yielded by the original halves and provided further the various intercorrelations are equivalent. Brownell ² has shown that in the case of objective tests constructed by an instructor for measuring the achievement of his students, different methods of forming the halves result in a

Wood, B. D. "Studies of Achievement Tests, Part III, Spearman-Brown Reliability Predictions," *Journal of Educational Psychology*, 17: 263-69, April, 1926.

Use of the formula is facilitated by the table prepared by Edgerton and Toops: Edgerton, H. A., and Toops, H. A. "A Table for Predicting the Validity and Reliability Coefficients of a Test When Lengthened," *Journal of Educational Research*, 18: 225-34, October, 1928.

¹ Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, p. 197.

² Brownell, W. A. "On the Accuracy with Which Reliability May Be Measured by Correlating Test Halves," *Journal of Experimental Education*, 1: 204-15, March, 1933.

wide range of values for $r_{\frac{1}{2} II}$. This condition is probably due, at least in part, to failure to satisfy the assumptions just noted. It is likely that they are not fully satisfied in more carefully constructed tests.

A reliability coefficient obtained from halves of a test and the application of the Spearman-Brown formula does not represent the same thing as one obtained from two applications of the test or the application of two forms of the test. Variations in the scores of individual pupils are due in part to differences in mind-set, effort, and the like. Such differences are probably eliminated in the case of two halves of the same test, but they are likely to be significant when there are two separate testings, especially when there is an interval of a day or more between them. Hence, the coefficient derived from the two halves of a test and the Spearman-Brown formula may be expected to be somewhat larger than would be obtained from a second administration of the test after an interval of at least a few hours.¹

The reliability of psychological questionnaires has been determined by the techniques used in determining the reliability of ordinary tests. Some investigators have administered their questionnaires twice to the same group of subjects and correlated the results. The weakness of this method is that the subjects may remember some of their initial responses, deliberately change them, thus reducing the reliability coefficient obtained. The use of equivalent forms is most desirable but it is often exceedingly difficult to construct equivalent forms of a psychological questionnaire.² Symonds, after summarizing

¹ For experimental verification, see Foran, T. G. "A Note on Methods of Measuring Reliability," *Journal of Educational Psychology*, 22: 383-87, May, 1931.

Jordan, R. C. "An Empirical Study on the Reliability Coefficient," *Journal of Educational Psychology*, 26: 416-26, September, 1935.

² Cady, following the suggestion of Kelley, rewrote his questionnaire so that each item required the opposite response. He does not consider this device as effective as the selection of a number of questions and the careful division of these into two forms to be administered at different sittings.

Cady, V. M. "The Estimation of Juvenile Incorrigibility," *Journal of Delinquency Monographs*, No. 2. Whittier, California: Whittier State School, 1923. 140 pp.

the reliability coefficients of a number of psychological questionnaires, concludes that these instruments compare favorably in reliability with ordinary tests of the recall, multiple-response, and true-false types.¹

The reliability of a battery of tests may be computed from the reliabilities and intercorrelations of the sub-tests by means of the formula for the correlation of sums.² If the sub-tests are approximately equivalent in reliability and the intercorrelations are high, the general form of the Spearman-Brown formula will give nearly the same result. When these conditions are not satisfied, the result obtained will be spuriously high.³

2. *Index of reliability.* The coefficient of reliability, r_{1I} , is an index of $X_1 - X_I$ rather than of $X_1 - X_\infty$. A measure of the latter may be obtained by calculating the *index of reliability* as indicated by the formula

$$r_{1\infty} = \sqrt{r_{1I}}$$

in which $r_{1\infty}$ represents the coefficient of correlation between a set of obtained scores and the corresponding set of theoretical true scores.

3. *Probable error of measurement.* The variable errors of measurement are the differences between the obtained scores (X_1) and the corresponding theoretical true scores (X_∞). The median deviation of these differences, which is called the *probable error of measurement*, is given by the formula⁴

$$PE_{1.\infty} = .6745\sigma_1 \sqrt{1 - r_{1I}}$$

The $PE_{1.\infty}$ may be interpreted with respect to the group to which the test was administered or with respect to any member

¹ Symonds, P. M. *Diagnosing Personality and Conduct*. New York: The Century Company, 1931, p. 168.

² See reference to Kelley on page 202.

³ Douglass, H. R., and Cozens, F. W. "On Formula for Estimating the Reliability of Test Batteries," *Journal of Educational Research*, 20: 369-77, May, 1929.

Handy, Urvan, and Lentz, T. F. "Item Value and Test Reliability," *Journal of Educational Psychology*, 25: 703-08, December, 1934.

⁴ For the derivation of this formula, see pages 132-33.

of that group. In the first case, the calculated $PE_{1.\infty}$ is the median deviation of the variable errors in the obtained test scores. When the interpretation of the calculated $PE_{1.\infty}$ is with reference to the score of a given pupil,¹ it is to be thought of as specifying the limit for which the chances are just fifty-fifty that the variable error will not be greater. For example, suppose $PE_{1.\infty} = 4.0$. Then for a given pupil the chances are fifty-fifty that his score involves a variable error not greater than 4.0. It is also possible to interpret the coefficient of reliability in terms of the per cent of pupils whose scores involve a variable error greater than a specified amount.²

The formula for the probable error of measurement affords a means for an interpretation of coefficients of reliability³ in terms of the corresponding median deviations of the variable errors of measurement. Unless the value of σ_1 is known, it will appear as a factor in this median deviation. Table IX gives for various values of r_{1I} the corresponding values of the probable error of measurement. It should be noted that unless σ_1 is relatively small, a reliability coefficient must approach 1.00 in order to indicate small variable errors of measurement.

Generalizing a measure of reliability. When a calculated coefficient of reliability, index of reliability, or probable error

¹ This application of the probable error of measurement implies the assumption that the variable errors are uncorrelated with the scores. This assumption is only approximately true. The larger errors appear to be found in the smaller scores and the smaller errors in the larger scores. However, the degree of correlation is not high and the application of the probable error of measurement to individual pupils seems to be justified, but the results should not be considered highly precise in the case of either very low scores or very high scores. See

Holzinger, K. J. "An Analysis of Errors in Mental Measurement," *Journal of Educational Psychology*, 14: 278-88, May, 1923.

² Herring, J. P. "The Verification of Group Examinations," *Journal of Educational Psychology*, 25: 596-602, November, 1924. In this article the formula used is not the one given on page 204. For calculations based upon this formula see

Huffaker, C. L. "The Reliability of Measurement by Group Tests of Mental Ability," *Journal of Educational Psychology*, 16: 493-95, October, 1925.

Herring, J. P. "Reply to Huffaker's Criticism," *Journal of Educational Psychology*, 16: 498-99, October, 1925.

³ A coefficient of reliability may also be interpreted in terms of the ratio of the variance (square of the standard deviation) of the true scores to the variance of the obtained scores. See Chapter XI.

of measurement is associated with a test as an index of the variable errors of measurement to be expected when it is administered to a group of pupils, generalization is implied. Representativeness of the group from which a determination of the index is made is a requirement for such generalization. As pointed out on page 110, a coefficient of correlation is affected by the variability or range of talent of the group. For a given measuring instrument, the coefficient of reliability calculated from a single grade group will be smaller than one calculated from a population including a sequence of grades. A coefficient of .60 calculated from a single grade group may be indicative of smaller variable errors of measurement than a coefficient of .90 calculated from a population including grades III to VIII. Hence, generalization of a coefficient of reliability must be restricted to populations of the same range of talent. Even in such cases, consideration of the causes of variable errors of measurement suggests that their magnitude may not be entirely stable. Hence, the generalization should not be considered highly dependable.

TABLE IX. VALUES OF COEFFICIENT OF RELIABILITY r_{II} AND CORRESPONDING VALUES OF PROBABLE ERROR OF MEASUREMENT $.6745\sigma_1\sqrt{1 - r_{II}}$

r_{II}	$.6745\sigma_1\sqrt{1 - r_{II}}$	r_{II}	$.6745\sigma_1\sqrt{1 - r_{II}}$
.50	.48 σ_1	.90	.21 σ_1
.55	.45 σ_1	.91	.20 σ_1
.60	.43 σ_1	.92	.19 σ_1
.65	.40 σ_1	.93	.18 σ_1
.70	.37 σ_1	.94	.17 σ_1
.75	.34 σ_1	.95	.15 σ_1
.80	.30 σ_1	.96	.13 σ_1
.82	.29 σ_1	.97	.12 σ_1
.84	.27 σ_1	.98	.10 σ_1
.86	.25 σ_1	.99	.07 σ_1
.88	.23 σ_1	1.00	.00 σ_1

The probable error of measurement is assumed to be independent of the population from which the determination is

made. On the basis of this assumption, a calculated probable error of measurement is commonly associated with a test as a measure of the magnitude of the variable errors of measurement to be expected in the scores yielded by it. This assumption is probably only approximated and hence a calculated probable error of measurement may not be a highly dependable index. Furthermore, when comparing tests with reference to their probable errors of measurement, it should be remembered that the significance of a probable error of measurement depends upon the range and magnitude of the scores yielded by the test. If the scores range from 125 to 200, a probable error of measurement of 5.0 has less practical significance than when the scores range only from 25 to 75.

Determining an index of the variable errors of validity to be expected in the scores yielded by a given test. Although we speak of determining the validity of a test, it should be remembered, as pointed out on page 174, that this characteristic of an instrument for measuring human abilities and traits is not stable. Hence, generalization from a determination for a given test is hazardous. Furthermore, it should be noted that the phrase, "validity of a test" has no reference to the systematic error of validity.

The determination of an index of the variable errors of validity requires criterion measures of the ability or trait specified by the function of the test.¹ In the case of a prognostic test, certain criterion measures are obtainable. For example, if the function of a test is to predict future scholastic success, which is defined as the school mark received in the field specified, the criterion measures are the school marks received by the pupils to whom the test is administered. If the test is designed to measure the same abilities or traits that are measured by an available test, the scores yielded by this instrument may be used as the criterion measures. For example, the Stanford Revision of the Binet Test

¹ For a critical discussion of the concept of validity, see Turney, A. H. "The Concept of Validity in Mental and Achievement Testing," *Journal of Educational Psychology*, 25: 81-95, February, 1934.

is believed to yield highly valid measures. Hence, mental ages obtained from its use are used as criteria for determining the validity of new intelligence tests. The more common case, however, is one in which the ability or trait to be measured is defined by a phrase such as "ability to read silently," "ability to solve arithmetical problems," "achievement in history," "language ability," "study habits," or "success in teaching" and no instrument is available for securing measures of it that are known to be highly valid. In such cases, testmakers have used school marks, teachers' estimates, and composite scores from a selected group of tests.¹ Obviously such measures, or combinations of them, are subject to criticism as criterion measures and hence determination of an index of the variable errors of validity to be expected in the scores yielded by a test is usually not very satisfactory. When criterion measures are available, the coefficient of correlation between them and the scores yielded by it, r_{1c} , is called the coefficient of validity.²

E. IMPROVEMENT OF TESTS

Revising a test to increase its reliability and validity. After a testmaker has determined the reliability and validity of his instrument, he may desire to investigate the possibilities of improving it in these respects. An obvious procedure is that of lengthening the test. The general form of the Spearman-Brown

¹ For partial summaries of procedures that have been employed in determining validity, see

Foran, T. G. "The Meaning and Measurement of Validity," *The Catholic University of America Educational Research Bulletins*, Vol. V, No. 7. Washington, D. C.: The Catholic University Press, September, 1930. 27 pp.

Kinney, L. B., and Eurich, A. C. "A Summary of Investigations Comparing Different Types of Tests," *School and Society*, 36: 540-44, October 22, 1932.

Jordan, A. M. "The Validation of Intelligence Tests," *Journal of Educational Psychology*, 14: 348-66, 414-28, September, October, 1923. This reference gives a comprehensive bibliography up to the date of its publication.

Lee, J. M., and Symonds, P. M. "New-Type or Objective Tests: A Summary of Recent Investigations," *Journal of Educational Psychology*, 24: 21-38, January, 1933.

Lincoln, E. A. "Studies of the Validity of the Dearborn General Intelligence Examinations," *Journal of Educational Psychology*, 19: 346-49, May, 1928.

² For the meaning of the coefficient of validity, see page 148. The interpretation is also considered in Chapter XI.

formula provides a means of estimating the reliability of a lengthened test.

$$r_{nn} = \frac{nr_{1I}}{1 + (n - 1)r_{1I}}$$

In this formula r_{nn} is the reliability coefficient of the lengthened test, and n is the number of times the length of the test has been increased.¹ Solving the equation for n , we have

$$n = \frac{r_{nn}(1 - r_{1I})}{r_{1I}(1 - r_{nn})}$$

When written in this form, the number of times the length of the test must be increased to secure a desired coefficient of reliability (r_{nn}) can easily be calculated.² On page 202 attention was called to the assumptions on which the derivation of the Spearman-Brown formula is based, and even when these conditions are satisfied the estimated reliability coefficient of a lengthened test should be considered a probable upper limit of the actual reliability.

Since the validity of a test is conditioned by its reliability, the validity will be increased by lengthening the test. The formula³ is

$$r_{nc} = \frac{nr_{1c}}{\sqrt{n + n(n - 1)r_{1I}}}$$

in which r_{nc} is the coefficient of validity of the lengthened test. It should be noted, however, that since

$$r_{1c} = r_{x_{\infty}c_{\infty}} \sqrt{r_{1I}} \sqrt{r_{cc}}$$

the upper limit of r_{nc} is $r_{x_{\infty}c_{\infty}} \sqrt{r_{cc}}$. Hence, if $r_{x_{\infty}c_{\infty}}$ does not ap-

¹ It should be noted that n may have fractional values. When $n = 2$, the formula simplifies to the form given on page 201.

² The following reference gives a useful table:

Edgerton, H. A., and Toops, H. A. "A Table for Predicting the Validity and Reliability Coefficients of a Test When Lengthened," *Journal of Educational Research*, 18: 225-34, October, 1928.

³ Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, p. 170.

proach 1.00, it will not be possible to make the test highly valid by lengthening it.

The reliability and validity of a test are affected by the difficulty of the test items,¹ administration time,² type of test exercises, method of scoring, directions for doing the exercises, and possibly by other factors. Hence, in revising a test to increase its reliability and validity, a testmaker should investigate the possibility of effecting improvement in ways in addition to increasing its length.

GENERAL REFERENCES AND SELECTED ILLUSTRATIONS

This bibliography is mainly a selected list of references dealing with the theory and technique in test construction. A few references of historical interest and a few texts describing tests available for use have been included. The references of the latter type will be helpful to an investigator who is seeking a test for a particular purpose. The bibliography by Hildreth is very complete and may be consulted in this connection.

AYRES, L. P. "A Measuring Scale for Ability in Spelling," *Russell Sage Foundation Bulletin E 139*. New York: Russell Sage Foundation, 1915. 56 pp.

Spelling scale based on one thousand words found to be of most common occurrence in a large amount of correspondence.

BINET, A., et SIMON, T. "Méthodes Nouvelles pour le Diagnostic du Niveau Intellectuel des Anormaux," *L'Année Psychologique*, 11: 191-244, 1905.

First individual intelligence test.

BOVARD, J. F., and COZENS, F. W. "Tests and Measurements in Physical Education, 1861-1925." *Oregon University Publications*, Physical Education Series, Vol. 1, No. 1. Eugene, Oregon: University of Oregon, 1926. 94 pp. See also: Bovard, J. F., and Cozens, F. W., *Tests and Measurements in Physical Education*. Philadelphia: W. B. Saunders Company, 1930. 364 pp.

BROWN, WILLIAM, and THOMSON, G. H. *Essentials of Mental Measurement*. London: Cambridge University Press, 1921. 216 pp.

¹ Thurstone, T. G. "The Difficulty of a Test and Its Diagnostic Value," *Journal of Educational Psychology*, 23: 335-43, May, 1932.

² Lindquist, E. F., and Cook, W. W. "Experimental Procedures in Test Evaluation," *Journal of Experimental Education*, 1: 163-85, March, 1933.

BUCKINGHAM, B. R. "Spelling Ability: Its Measurement and Distribution," *Teachers College, Columbia University Contributions to Education*, No. 59. New York: Bureau of Publications, Teachers College, Columbia University, 1913. 116 pp.

First example of a measuring instrument in which difficulty values were determined from per cents of correct response and in which the items were arranged in order of increasing difficulty.

COURTIS, S. A., et al. "The Measurement of Educational Products," *Seventeenth Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1918. 192 pp.

An excellent source for information respecting the status of the measurement movement in 1918.

DEARBORN, W. F. *Intelligence Tests*. Boston: Houghton Mifflin Company, 1928. 336 pp.

FREEMAN, F. N. *Mental Tests*. Boston: Houghton Mifflin Company, 1926. 503 pp.

FRYER, DOUGLAS. *The Measurement of Interests in Relation to Human Adjustment*. New York: Henry Holt and Company, 1931. 488 pp.

HARTSHORNE, HUGH, and MAY, M. A. *Studies in Deceit*. New York: The Macmillan Company, 1928. 414 and 306 pp. (Book One and Book Two are bound in the same volume which has the title "Studies in the Nature of Character.")

Pioneer research in the measurement of character traits.

HILDRETH, G. H. *A Bibliography of Mental Tests and Rating Scales*. New York: The Psychological Corporation, 1933. 242 pp. This compilation has been supplemented by Buros, O. K. "Educational, Psychological, and Personality Tests of 1933 and 1934," *Studies in Education, Rutgers University Bulletin*, Vol. XI, No. 11. New Brunswick, New Jersey: Rutgers University, 1935. 44 pp.

HILLEGAS, M. B. "A Scale for the Measurement of Quality in English Composition by Young People," *Teachers College Record*, 13: 331-84, September, 1912.

First English composition scale.

HULL, C. L. *Aptitude Testing*. Yonkers-on-Hudson, New York: World Book Company, 1928. 535 pp.

KELLEY, T. L. *Interpretation of Educational Measurements*. Yonkers-on-Hudson, New York: World Book Company, 1927. 363 pp.

KWALWASSER, JACOB. *Tests and Measurement in Music*. Boston: C. C. Birchard and Company, 1927. 146 pp.

MCCALL, W. A. "A Proposed Uniform Method of Scale Construction," *Teachers College Record*, 22: 31-52, January, 1921.

Proposal of the T-score technique.

MONROE, W. S. *An Introduction to the Theory of Educational Measurements*. Boston: Houghton Mifflin Company, 1923. 364 pp.

The first comprehensive treatment of the theory of educational measurements.

ODELL, C. W. *Educational Measurement in High School*. New York: The Century Company, 1930. 641 pp.

ODELL, C. W. *Traditional Examinations and New Type Tests*. New York: The Century Company, 1928. 469 pp.

OTIS, A. S. "An Absolute Point Scale for the Group Measurement of Intelligence," *Journal of Educational Psychology*, 9: 239-61, 333-48, May, June, 1918.

First group intelligence test.

PETERSON, JOSEPH. *Early Conceptions and Tests of Intelligence*. Yonkers-on-Hudson, New York: The World Book Company, 1925. 320 pp.

Early history of intelligence testing.

PINTNER, RUDOLF. *Intelligence Testing: Methods and Results*. New York: Henry Holt and Company, 1931. 555 pp.

RUCH, G. M. *The Objective or New Type Examination*. Chicago: Scott, Foresman and Company, 1929. 478 pp.

RUCH, G. M., and STODDARD, G. D. *Tests and Measurement in High School Instruction*. Yonkers-on-Hudson, New York: World Book Company, 1927. 381 pp.

SYMONDS, P. M. *Diagnosing Personality and Conduct*. New York: The Century Company, 1931. 602 pp.

SYMONDS, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927. 588 pp.

TERMAN, L. M. *The Intelligence of School Children*. Boston: Houghton Mifflin Company, 1919. 317 pp.

Supplements the earlier volume *Measurement of Intelligence* (1916) which is largely devoted to the actual administration of the Stanford Revision.

TERMAN, L. M., and CHILDS, H. G. "A Tentative Revision and Extension of the Binet-Simon Measuring Scale of Intelligence," *Journal of Educa-*

CONSTRUCTING MEASURING INSTRUMENTS 213

tional Psychology, 3: 61-74, 113-43, 198-208, 277-89; February, March, April, May, 1912.

First appearance of the Stanford Revision of the Binet-Simon Scale.

THORNDIKE, E. L. *An Introduction to the Theory of Mental and Social Measurements*. New York: Teachers College, Columbia University, 1904. 277 pp. (Revised edition, 1913.)

The pioneer book in its field.

THORNDIKE, E. L. "Handwriting," *Teachers College Record*, 11: 1-93, March, 1910.

First quality scale.

THORNDIKE, E. L., et al. *The Measurement of Intelligence*. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 616 pp.

WILSON, G. M. "The Purpose of a Standardized Test in Spelling," *Journal of Educational Research*, 20: 319-26, December, 1929.

Discusses a basic problem in test standardization.

WILSON, G. M., and HOKE, K. J. *How to Measure*. New York: The Macmillan Company, 1928. 597 pp. (Revised and enlarged.)

WOOD, B. D. *Measurement in Higher Education*. Yonkers-on-Hudson, New York: World Book Company, 1923. 337 pp.

YERKES, R. M. (Editor). "Psychological Examining in the United States Army," *Memoirs of the National Academy of Sciences*, Vol. 15, Washington: Government Printing Office, 1921. 890 pp.

Describes the use of the Army Alpha and Army Beta in 1917-18.

CHAPTER VIII

STUDYING CURRENT CONDITIONS OR PRACTICES

The general character of survey investigations. The general character of studies of current conditions and practices is illustrated by surveys of school systems¹ whose findings include such items as per capita expenditures for educational purposes, average salary of teachers, average pupil achievement as revealed by certain tests, average size of class, facts pertaining to school buildings, years of training and experience of teachers, and age-grade status of pupils. Frequently a survey is less comprehensive, but the general purpose is to reveal the status of what is studied. The following titles are indicative of typical purposes.²

- A Tentative Inventory of the Habits of Children from Two to Four Years of Age (2)
- The Vocabulary of American History (5)
- Mistakes Which Pupils Make in Spelling (10)
- The Status of the Superintendent (13)
- The Social Composition of Boards of Education (16)
- The Duties of the Elementary-School Principal (20)
- Educational Magazines Read by Five Hundred Elementary School Principals and Classroom Teachers (25)
- A Survey of the Requirements for the Doctor of Philosophy in Education (34)
- Subject Combinations in the Programs of Teachers in Small Secondary Schools in New York State (35)

¹ For a brief account of the development of the survey movement see Caswell, H. L. "City School Surveys," *Teachers College, Columbia University Contributions to Education*, No. 358. New York: Bureau of Publications, Teachers College, Columbia University, 1929, Chapter II.

It is interesting to learn that at the invitation of the governor of Rhode Island, Henry Barnard made a survey of the public schools of that state in 1845. There was also a survey of school achievement in Boston during the same year.

² The numbers in parentheses following the titles refer to the illustrative bibliography at the end of the chapter.

The Training of Modern Foreign Language Teachers in the United States (45)

The Diversity of High School Students' Programs (49)

Sometimes the purpose is extended to include a comparison of the status of two or more populations as indicated by the following titles.

A Comparative Study of White and Colored Pupils in a Southern School System (23)

A Comparison of the Achievement of Eighth-Grade Pupils in Rural Schools and in Graded Schools (31)

A Comparative Study of the Physical Growth of Dull Children (51)

Inquiries of the survey type range from elaborate studies of city or state school systems, and even national inquiries, to simple studies pertaining to a limited area. Occasionally a survey is narrowed down to an intensive study of a single pupil or a small group of pupils studied separately. Such surveys are called "case studies." Facts pertaining to current conditions and practices frequently become more meaningful when they are compared with information respecting corresponding conditions and practices of the past. Such comparisons may reveal trends which are useful in predicting the future. When the trend studied is that of the growth of a human trait, the research is frequently designated as "genetic." Sometimes the data collected in a survey are utilized as a basis for investigating the existence of relationships. Such studies transcend the types of research to be considered in this chapter except when the techniques employed are relatively simple.¹

The definition of survey problems. Frequently the problem of a survey is first conceived of in general terms, but in order to serve as a guide for the subsequent phases of the investigation, it must be defined. This definition should include the specification of the scope of the survey, the specific questions for which answers are to be sought, and the meaning of the technical terms employed. As the investigator engages in the later stages of his

¹ Applications of correlation analysis in the study of relationships are considered in Chapter XI.

research, he may decide to modify some of the questions listed or even to omit certain ones. Sometimes he may become interested in increasing the scope of his study. The possibility of such changes, however, does not lessen the desirability of formulating an effective definition as a first step. When a survey has been carefully planned in advance, its later stages, with the exception of interpretation, tend to become routine in character—an important consideration in large-scale investigations where most of the labor must be done by clerical workers.

Adequate definition of terms is especially important when a specified trait or characteristic cannot be measured directly. For example, if the survey is to determine the "average teaching load" of the high schools within a certain area, measures of this characteristic cannot be obtained directly but must be computed from such basic items as number of students, number of teachers, and hours per week spent in classrooms. Since no standard method has been established for computing the average teaching load of a high school, an investigator must formulate a definition in terms of the calculations to be made. If this is not done in planning the survey, he may find that he has failed to collect some of the required data.

A. TECHNIQUES OF SURVEYS

The data and their collection.¹ The data of surveys vary. They include facts pertaining to such conditions and practices as school costs; educational legislation; duties or activities of school people; practices with respect to state control of education; overlapping of courses; vocabularies of textbooks; vocational opportunities of high school graduates; achievement, intelligence, and other traits of pupils; socio-economic status of

¹ The interested reader may consult the following for information relative to practices in collecting survey data:

Caswell, H. L. "Survey Techniques," *Educational Administration and Supervision*, 19: 431-41, September, 1933.

Davis, E. C. "Methods and Techniques Used in Surveying Health and Physical Education in City Schools," *Teachers College, Columbia University Contributions to Education*, No. 515. New York: Bureau of Publications, Teachers College, Columbia University, 1932. 162 pp.

homes; current theories in education; recognized objectives; numbers and titles of books in high school libraries; styles of architecture of school buildings; practices in scheduling recitations; and provisions of teachers' contracts. The sources of data include published materials such as books, periodicals, and monographs in the field of education and unpublished materials such as school records. They include school people such as superintendents, principals, and teachers from whom data may be collected by correspondence, by interview, or by observation. They include pupils from whom the data may be collected by the administration of tests or questionnaires, or by observation and interview. School buildings, their equipment, and their environment may also be included as sources of data collected by observation.

The basic techniques employed in collecting data were considered in Chapter III, but a few points may be emphasized here. When data are copied from records or published reports, the investigator should inquire into their accuracy and make certain of the precise meaning of the label of each item. Interviewers and observers should receive training before beginning to collect data for a survey. Training is also desirable when analysis is being employed. If possible, at least a portion of the analysis should be checked by a second worker. Questionnaires should be constructed with care. If the investigator is inexperienced, he should secure the criticism of competent persons and if possible try out the questionnaire by submitting it to a representative group not included in the survey. In selecting a test, its function should be noted. If it is necessary to construct a test, the advice of competent persons should be sought and an attempt should be made to determine the function of the completed instrument. In administering a test, the instructions should be followed unless deviations are considered desirable. When this is the case, changes in the instructions should be determined and reduced to writing.

Securing representativeness of data in survey investigations. When the scope of a survey is comprehensive, it may not be

feasible to collect data for the entire population designated by the problem. In such cases some process of sampling must be resorted to. A sample may be regarded as satisfactory to the extent that the data obtained are representative of the specified population or area. Hence, the investigator faces the problem of obtaining a highly representative sample.

Occasionally a method of random sampling may be employed. A random sample is not representative except by chance, but the probable deviation from representativeness may be calculated. When a method of random sampling is not employed, the investigator may be able to secure a highly representative sample. If the larger population or universe is stratified, a selection should be made from each stratum. The size of the samples should be proportional to the size of the respective strata. For example, if data are to be collected from school systems of varying size and there are many more small systems than large ones, the number of small systems investigated should be proportionally greater than the number of large ones. However, if the small systems are more homogeneous in character than the large ones, the per cent of small systems investigated will not need to be as large as the relative numbers of large and small systems would indicate. These principles can be applied in the collection of various types of data. For example, when textbooks are analyzed and sampling must be employed, the size of the samples should be proportional to the space given to important topics, lengths of the texts, number of texts in different subjects of the problem, and so on, according to the requirements of the problem. A text of limited range of vocabulary requires a relatively smaller sample than a text whose vocabulary range is extensive.

When the data are collected by means of a questionnaire, the selection is determined by influences that the investigator can control only indirectly. When observation or interview is employed, accessibility is likely to be a determining factor. In such cases determination of the degree of representativeness is an important subordinate problem. There is no single technique

for accomplishing this. When there is justification for expecting that the frequency distributions of a representative sample will approach the normal shape, this criterion may be applied. The construction of a frequency polygon or histogram will be sufficient in many cases to indicate whether or not an approximately normal distribution has been obtained. When a more precise determination is desired, Pearson's χ^2 test may be applied.¹ Thomson and Pintner have suggested that "one criterion for the unselected nature of any group of children is that the coefficient of correlation of age and mental age in such a group should be approximately equal to the ratio of the coefficients of variability of chronological and mental ages."² It should be noted, however, that in certain cases a normal distribution should not be expected. For example, a representative sample of "gifted" children should yield a distribution of intelligence quotients far different from a normal one.

The representativeness of newly collected survey data may frequently be tested by comparison with similar data reported in such published sources as the *Federal Census*, the *Biennial Surveys of Education* of the United States Office of Education, and reports of national, state, or city surveys. In making such evaluations of the representativeness of new data, the meaning and comparability of units should be noted. The possibility of changed conditions since the publication of the criterion should also be considered.

In the study of trends, smaller samples may be adequately representative of conditions in the earlier years, if the passage of time has resulted in increases of populations and in their heterogeneity. In considering the representativeness of pupil populations over a period of years, the operation of selection must be recognized as a significant factor. Thirty years ago the high

¹ See Chapter IV, page 79.

² Thomson, G. H., and Pintner, Rudolph. "Spurious Correlation and Relationship between Tests," *Journal of Educational Psychology*, 15: 433-44, October, 1924. For an application of this test, but not in a survey study, see Heilman, J. D. "Factors Determining Achievement and Grade Location," *Journal of Genetic Psychology*, 36: 439-40, September, 1929. See page 234 for the formula for the coefficient of variability.

school population was highly select in comparison with the children of high school age of that time. At present, approximately one child of high school age out of every two children is in high school. If a group of school children is studied over a period of years, elimination from school will decrease the extent to which the group is representative of children "in general." In studying trends in the development of traits, different groups are frequently selected from the different grade levels. These groups may not be comparable in representativeness to a single group studied over a period of years.

Computing derived measures.¹ A ratio, frequently expressed as a per cent, is a very common derived measure. For example, one may wish to express the score of a pupil on a spelling test in terms of the per cent of words he has spelled correctly. It may be desired to determine what per cent of educational expenditures is for teachers' salaries. A ratio may be computed for any convenient base. "Pupils per thousand population," "pupils per teacher," "expenditures per pupil in average daily attendance," "school indebtedness per capita," and "language errors per thousand words of written expression" are illustrations. In his appraisal of secondary school commercial education, W. R. Odell calculated the ratio of public school enrollments in certain commercial vocational subjects to the number of workers in the vocations as shown by the Census of 1930.²

The calculation of ratios is simple, but attention should be given to the basic data, especially when comparisons are to be made. If two cities are to be compared with respect to per cent of educational expenditures devoted to teachers' salaries, the term "educational expenditures" should have the same meaning for both cities, i.e., the same items of expense should be included. It should not include in one case "interest on bonds" and exclude it in the other. Furthermore, the term "teacher" should have the same meaning in both systems. It should not

¹ For the calculation of age scores, T-scores, percentile scores, and the like, consult the Index of this volume.

² Odell, W. R. "An Appraisal of Secondary School Commercial Education," *Teachers College Record*, 34: 43-52, October, 1932.

mean in one case merely classroom teachers and include in the other case elementary principals who devote some of their time to teaching. Hence, the measures from which a ratio is to be calculated should be defined with precision, and if comparisons are to be made, the investigator should make certain that the measures conform to these definitions in each population unit.

A sum or average may be desired as a derived measure. For example, if several achievement tests have been administered, it may be desired to obtain a composite measure of achievement. Since the scores yielded by different tests are seldom expressed in terms of equivalent units and from a common zero point, transformation of the obtained scores to a common scale is a logical preliminary step.¹ The transformed measures may then be combined as desired. Frequently a special formula can be derived which will simplify the total process. If criterion measures are available, multiple regression may be employed as a means of determining such a formula.

Raw data are sometimes combined to form complex index measures. On the basis of the number of school districts, the average daily attendance in elementary schools, the average daily attendance in high schools, the density of rural school populations, and expenditures for transportation in rural districts, Mort² has calculated indices of educational need by means of a formula derived through the use of techniques of curve-fitting. Burns³ employed a similar technique in calculating an index of transportation need. An index measure of school building utilization has been devised by Morphet.⁴ For

¹ See pages 82-83.

² Mort, P. R. "The Measurement of Educational Need. A Basis for Distributing State Aid," *Teachers College, Columbia University Contributions to Education*, No. 150. New York: Bureau of Publications, Teachers College, Columbia University, 1924. 85 pp.

³ Burns, R. L. "Measurement of the Need for Transporting Pupils," *Teachers College, Columbia University Contributions to Education*, No. 289. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 61 pp.

⁴ Morphet, E. L. "The Measurement and Interpretation of School Building Utilization," *Teachers College, Columbia University Contributions to Education*, No. 264. New York: Bureau of Publications, Teachers College, Columbia University, 1921, pp. 21 f.

use at the college level, Reeves and Russell¹ proposed as a "weighted index of teaching load" the sum of the ratio of an instructor's "teaching hours" to the average for his institution, the ratio of his "preparation hours" to the institutional average, and two times the ratio of his "student hours" to the institutional average. An index measure for textbooks has been proposed by Patty and Painter² and a "student ability index" by Banker.³ In 1920, Ayres⁴ proposed as an index measure for state school systems a composite of ten different elements of which five are measures of the amount of education received by children and five are measures of the expenditures made to purchase this education. Other indices are to be found in studies reported by Burgess,⁵ Norton,⁶ and Clark.⁷

¹ Reeves, F. W., and Russell, J. D. *College Organization and Administration*. Indianapolis, Indiana: Board of Education, Disciples of Christ, 1929, pp. 176 f.

For another index measure of teaching load, see Douglass, H. R. *Organization and Administration of Secondary Schools*. Boston: Ginn and Company, 1932, pp. 114 f.

² Patty, W. W., and Painter, W. I. "A Technique for Measuring the Vocabulary Burden of Textbooks," *Journal of Educational Research*, 24: 127-34, September, 1931.

³ Banker, H. J. "A Student's Ability Index from Teacher's Marks," *Journal of Educational Research*, 17: 357-64, May, 1928.

Banker, H. J. "The Practical Application of the Student's Ability Index," *Journal of Educational Research*, 18: 282-89, November, 1928.

Banker, H. J. "The Student's Ability in Higher Educational Institutions," *Journal of Educational Research*, 26: 276-83, December, 1932.

⁴ Ayres, L. P. *An Index Number for State School Systems*. New York: Russell Sage Foundation, 1920. 70 pp.

Ayres published an earlier study in which the states were ranked according to ten characteristics. Ayres, L. P. *A Comparative Study of Education in the 48 States*. New York: Russell Sage Foundation, 1912.

See also Phillips, F. M. *Educational Ranking of States by Two Methods*. Milwaukee: Bruce Publishing Company, 1925.

Phillips, F. M. "Educational Rank of States, 1930," *American School Board Journal*, 84: 25-29, 29-30, 37-39, 39-40, February, March, April, and May, 1932.

⁵ Burgess, W. R. *Trends of School Costs*. New York: Russell Sage Foundation, 1924. 142 pp.

⁶ Norton, J. K. "The Ability of the States to Support Education," *Research Bulletin of the National Education Association*, Vol. 4, No. 1-2. Washington: National Education Association, 1926. 88 pp.

⁷ Clark, H. F. "The Effect of Population upon Ability to Support Education," *Journal of Educational Research*, 14: 336-39, December, 1926.

A more complete discussion is given in Clark, H. F. "The Effect of Population upon Ability to Support Education," *Bulletin of the School of Education*,

Clark ¹ has discussed the uses of index numbers in education. He stresses their importance in four different phases of school administration: (1) teachers' salaries and costs of living; (2) school buildings; (3) school bonds; (4) instructional supplies. The procedures, by means of which such index measures are calculated, represent attempts to give an appropriate weight to certain aspects of a complex condition or phenomena. Some of the resulting indices are reasonably satisfactory measures, but, in general, an investigator who calculates indices should bear in mind the possibility that the measures obtained will involve relatively large variable errors.

An investigator may desire the ranking of the members or units of the population being surveyed rather than their absolute measures. For example, he may desire the ranks of the pupils of a class rather than their scores on a test. If none of the units, or individuals, have the same raw measures, numbering from the highest to the lowest on the scale gives their respective ranks. If two or more individuals have the same score, or measure, an average rank is given as in the following example.

RAW MEASURE	RANK
24	1
20	2.5
20	2.5
17	4
15	5
14	7
14	7
14	7
10	9
8	11.5
8	11.5
8	11.5
8	11.5
7	14
5	15

Indiana University, Vol. 2, No. 1. Bloomington, Indiana: Indiana University, 1925. 29 pp.

¹ Clark, H. F. "Index Numbers in Educational Work," *Teachers College Record*, 30: 453-60. February, 1929.

The interested reader should consult the articles on school-bond prices, which

It should be noted that the final rank equals the number of cases, unless two or more raw measures have the same value at the bottom of the series.

Another method of assigning ranks is that of computing the percentile rank of each measure in the series, or of estimating such ranks by comparison with percentile points as calculated or as located on a percentile curve.¹ In a simple series of measures the percentile rank of a given one may be easily computed by calculating the per cent of measures below that one in the series. It should be noted that the order of percentile ranks is the reverse of ordinary ones. An ordinary rank of "1" is the highest on the scale, a percentile rank of "1" means that only one per cent of the measures are below that measure in the series.

If the raw data are in the form of rankings, transformation into amount scores may be effected providing the assumption is made that the distribution of the trait or characteristic approaches the normal shape in the population ranked.² When each of the members of the population has been ranked by several judges, the technique employed in constructing a quality scale may be used for obtaining average amount scores.³

When the members of a population have been ranked with reference to several traits or characteristics or have been ranked with reference to a given trait or characteristic by several have been published by Clark since January, 1928, in the *American School Board Journal*.

For a general treatment of index numbers, consult:

Chaddock, R. E. *Principles and Methods of Statistics*. Boston: Houghton Mifflin Company, 1925, pp. 175-206.

Fisher, Irving. *The Making of Index Numbers*. Boston: Houghton Mifflin Company, 1923. 526 pp.

Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, pp. 331-47.

¹ See Chapter IV, page 84.

² For a description of the method and a table to facilitate the transformation, see Hull, C. L. "The Computation of Pearson's r from Ranked Data," *Journal of Applied Psychology*, 6: 385-90, December, 1922.

³ For a description of the procedure, see Monroe, W. S. *Introduction to the Theory of Educational Measurements*. Boston: Houghton Mifflin Company, 1923, pp. 133-44.

judges, the obtained rankings may be combined by computing the mean or median of the assigned ranks. One may also total the ranks assigned to each member of a population and use these sums as a basis for an average ranking. If some of the series of rankings are incomplete, an average rank may be computed by transforming each set of rankings into amount scores and then computing the average score. From these average scores, an average ranking may easily be obtained.¹

Classification of data. If only a single frequency distribution is desired, the classification of data is usually a simple matter, but frequently separate tabulations are desired for certain sub-populations. For example, if the high schools of a state are being surveyed with reference to the number of pupils per teacher, the investigator may desire separate tabulations for schools enrolling less than one hundred pupils, schools enrolling one hundred up to one hundred fifty pupils, schools enrolling one hundred fifty up to two hundred pupils, and so on. He may desire also separate tabulations for geographical divisions of the state. In elaborate surveys the determination of the sub-populations for which separate tabulations are to be made, may require careful consideration. Survey findings are more meaningful when they are for populations that are relatively homogeneous with respect to significant characteristics, but a large number of sub-populations makes the interpretation more difficult. In determining sub-populations, an investigator should be guided by the definition of his problem. Sometimes the determination cannot be accomplished until after some analysis of the data has been made. The classification finally adopted will often be the result of the trial of a series of classifications. Sometimes a careful study of the previous researches in the field of the survey will suggest an effective classification. When the findings are to be compared with surveys previously made, the classifications should be similar. In general, it is wise to make the sub-popula-

¹ For a description of the procedure and other methods, see Garrett, H. E. "An Empirical Study of the Various Methods of Combining Incomplete Order of Merit Ratings," *Journal of Educational Psychology*, 15: 157-71, March, 1924.

tions rather highly homogeneous. After the data have been tabulated, certain sub-populations may be combined if it appears that the more elaborate analysis is not useful.

When the sub-populations are determined by variations in a single characteristic, the data may be tabulated in the form of a correlation table. The age-grade table is an illustration. If salaries of high school teachers are being tabulated, sub-classification with respect to both sex and size of school may be accomplished by drawing up a form for a correlation table with appropriate salary and size of school intervals and then dividing each of the columns to provide separate tabulations for the sexes. This form of table, however, becomes too complex when the number of subdivisions of a column is greater than two or three.

When the data are quantitative measures such as test scores, teachers' salaries, or pupil-teacher ratios, the classification and tabulation may be combined in a single process. But in handling other types of data, it may be desirable to effect a classification before the tabulation is attempted. For example, consider a survey of the participation of college students in extra-curricular activities in which the raw data include the names of the activities each student participates in. Before attempting to tabulate this information, its classification should be decided upon and the one or more classes in which each student falls should be marked on his record. For a given activity, a student may be given one of the following classifications: "not participating," "participating in the given activity only," "participating in the given activity and one other," "participating in the given activity and two others," and so on. The final classification in this series might be "participation in the given activity and four or more others." Each student would be given his classification for each activity considered in the survey. A more detailed classification would result from recognizing each combination of activities such as football and basketball, football and track, football and baseball, and the like.

In surveys such as those of the language errors of pupils the

classification is made on the basis of certain criteria which frequently must be formulated by the investigator. The problem is similar to that involved in collecting data by means of analysis which was discussed in Chapter III. The particular criteria that should be recognized in a given case depends upon one's purpose. An investigator should familiarize himself with a number of similar studies. It may be desirable to seek the advice of experienced persons.

Where data are obtained by means of questionnaires or interview schedules, the returns may be partially incomplete. Hence, it is usually advisable to recognize as a sub-class "no data" or "not given." In some cases there may be justification for combining "no data" with "none." For example, if teachers are asked "How many courses in the history of education have you had?" "No data" or "not given" may be taken as "none." It is a better practice, however, to restrict each frequency distribution to the number reporting specific information under the caption of the distribution. This practice will cause some variation in the numbers of cases represented in each distribution. If the number of cases is markedly different from that of the total population, the data relating to the given characteristic are probably less representative than where the per cent of response approaches one hundred. It may be advisable to exclude incomplete blanks from the tabulation. In any case, a high per cent of incomplete responses should be recognized as a data fault.

The mechanics of tabulating survey data. The labor of tabulating survey data may be materially lessened and the accuracy of the results increased by giving attention to the form in which the data are collected or recorded, especially when the survey is elaborate. Usually it is advantageous to employ individual data cards. For example, if the survey is one in which there are several items of information such as chronological age, sex, intelligence test score, average school mark, and vocational interests, the data for each student should be recorded in a separate card. A given item of information should

appear in the same position on each card. When the number of items is large, the card should be ruled so that a space will be provided for each item. In many cases this ruling may be conveniently accomplished by mimeographing, but in a comprehensive survey the record card should be printed. The spaces of the form may be labeled to indicate the items of information recorded in them. A convenient plan of labeling is to number the spaces in serial order. A general individual data card may be made by dividing the area into squares or rectangles of convenient size and then numbering them serially. Such a card may be used in various types of surveys. It may be advantageous to use cards of two or more colors. For example, in a survey of school children, white cards may be used for boys and buff ones for girls.

When all of the data of a survey are collected by means of a single questionnaire, the returned sheets or booklets constitute individual records and the desired tabulations can be conveniently made from them. Frequently the tabulation may be facilitated by anticipating this procedure in planning the form of the question blank. Similar statements may be made relative to the blank upon which data are recorded when they are copied from records or secured by other techniques.

Sometimes the process of tabulation may be materially facilitated by translating the raw data into a code in which the variations in an item of information are represented by numbers. In determining this code, it is necessary to anticipate the class intervals or rubrics in the subsequent tabulation. When these have been determined, they may be numbered 0, 1, 2, 3, . . . A given datum is then assigned the number corresponding to the class interval or rubric in which it falls. The following are illustrations of coding.¹

¹ For other illustrations and a discussion of the principles of coding, see Toops, H. A. "Some Considerations Relative to the Standardization of Certain Procedures in Educational Research," *Journal of Experimental Education*, 1: 229-38, March, 1933.

A more comprehensive treatment is given by Baehne, G. W. *Practical Applications of the Punched Card Method in Colleges and Universities*. New York: Columbia University Press, 1935. 442 pp.

CODE NUMBERS	AGES OF ADULTS IN YEARS	CODE NUMBERS	CURRICULA
0	No data	0	Agriculture
1	20-24	1	Athletic coaching
2	25-29	2	Biological science
3	30-34	3	Chemistry
4	35-39	4	Commerce
5	40-44	5	Education
6	45-49	6	Engineering
7	50-54	7	Law and pre-legal
8	55-59	8	Liberal arts
9	60 and above	9	Home economics
		10	Physics
		11	Social science

CODE NUMBERS	PARTICIPATION IN FOOTBALL AND BASKETBALL
0	No football, or no data
1	Football only
2	Football and one other activity
3	Football and two other activities
4	Football and three or more other activities
5	Basketball only
6	Basketball and one other activity
7	Basketball and two other activities
8	Basketball and three or more other activities
9	No basketball, or no data

Coding is merely a classification of data plus a technique for designating the results. Hence, when coding is employed, the classification of data is separated from the process of tabulation. This is desirable in elaborate surveys, especially when the variations to be recorded are of the type shown in the last of the preceding illustrations. Incidentally, it should be noted that the employment of code numbers is likely to reduce greatly the space required for recording the data on individual cards.

In surveys of large populations, the labor of tabulating may be greatly lessened by employing machines. The basis of mechanical tabulation is an individual record card on which the data in terms of code numbers are recorded by the position of punched holes. After the cards have been correctly punched, the sorting and counting is accomplished by machines.¹

¹ For a brief description of the use of mechanical tabulation in handling school

SERIAL NUMBER OF CHILD	SEX		HIGH SCHOOL CLASS				AGE							INTELLIGENCE QUOTIENT					
	M	F	1	2	3	4	13 and under	14	15	16	17	18	19 and over	Be- low 80	80-	90-	110-	120-	140 and above
1	1		1			1	1					1				1		1	
2	1		1					1								1			
3		1														1			
4	1			1					1							1			
5		1	1				1									1			
6		1			1			1		1						1			
7		1		1							1					1			
8	1														1				
9	1		1						1										
10		1		1					1							1			
11	1		1					1								1			
12		1	1					1									1		
13		1				1													
14	1		1						1										
15		1	1												1				
16		1		1						1						1			
17		1			1						1								
18	1					1							1						
19		1								1									
20		1		1					1							1			
21	1			1					1							1			
22	1				1											1			
23	1							1								1			
24	1			1												1			
25		1														1			
Total	12	13	9	7	5	4	1	7	7	4	3	2	1	0	3	15	6	1	0

Fig. 2. A tabulation sheet for four items of information, sex, high school class, age and IQ.

Sometimes an investigator records his data in the form illustrated in Figure 2. The various items of information for a given member of the population are recorded on a single line. In Figure 2 the raw data have been classified and the information is recorded by means of check marks in the appropriate columns. The totals of the checks in the columns give the frequency distributions for the four items of information. When the population is greater than the number of lines on the sheet, the subtotals may be carried forward. This type of record sheet is inconvenient when separate tabulations of an item are desired for the categories of another item. For example, if one wishes to obtain the distribution of intelligence quotients of sophomore girls, one must identify the checks that refer to these students and then make a tabulation on a separate sheet. The labor of doing this may be lessened to some extent by making a preliminary sorting before recording the data so that the information relative to sophomore girls will appear on consecutive lines. There is, however, a limit to such preliminary sorting. Hence, the type of record sheet shown in Figure 2 is seldom to be recommended.

Calculations from tabulations. The results of elaborate manipulations of data are usually difficult to interpret. Hence, in making calculations from tabulations, the investigator should use as simple techniques as are consistent with the demands of his problem and the nature of his data. Elaborate procedures resulting in precisely expressed statistical measures frequently lend a false air of dependability to the results obtained. Uncritical readers of a report of research are likely to accept at "face value" statistics whose calculation is not understood by them. The investigator should not, however, over-simplify the statistical treatment of his data. For example, an average gives

data, see Rugg, H. O. *Statistical Methods Applied to Education*. Boston: Houghton Mifflin Company, 1917, pp. 66 f.

Information concerning the various types of tabulating machines may be secured by addressing the International Business Machines Corporation, Tabulating Machine Division, 270 Broadway, New York City. This company maintains offices in a number of other cities.

only the central tendency of the conditions or practices. It does not reveal variations from this central tendency and they may be as significant as the average status. Measures of skewness, coefficients of variability, and other infrequently used statistics may be helpful in indicating the meaning of survey data.

The techniques for calculating means, medians, modes, quartile and percentile points, and measures of variability were described in Chapter IV. When the distribution does not approximate the normal shape or when the number of cases is small, care should be exercised in interpreting these statistics as summary measures. For example, if the distribution exhibits extreme skewness, the mean may be misleading as a central tendency. When the number of cases is small, percentile points and even quartile points should not be considered precise determinations. In some cases the calculation of these statistics may not be advisable.

In converting the frequencies of a distribution into per cents, a convenient procedure is to calculate the reciprocal of N , or the base. The desired per cents are then obtained by using this quotient as a multiplier of the frequencies. The products may be read from an appropriate table or obtained by employing a calculating machine. In the latter case the reciprocal is set in the keyboard and, using the ordinary procedure of multiplication, the frequencies are made to appear successively in the upper right dial of the machine and just below each frequency, as it appears, is the corresponding per cent. The summation of the per cents should equal 100., or 100.00 if two decimal places are carried, but, because of rounding off of decimal places, it may not quite equal this figure. Chaddock ¹ advocates, on logical grounds, the reporting of the sum as 100.00 even in cases where it is not precisely that amount. Adjustment of certain per cents so that the total exactly equals 100 reduces their accuracy and is undesirable when the adjusted per cents are given individual interpretations.

¹ Chaddock, R. E. *Principles and Methods of Statistics*. Boston: Houghton Mifflin and Company, 1925, p. 412.

When there is only a single distribution, the base to be used in calculating the per cents is usually apparent, but when there are several distributions and comparisons to be made, the determination of the bases to be used may require careful consideration in order to obtain per cents from which the desired interpretation may be derived. If no report or usable information has been obtained from some members of the sub-populations of a survey, the investigator must choose between the "total number in a sub-population" and the "number reporting usable information." Usually the latter is preferred as a base in calculating per cents. Another case that requires attention is when certain individuals or units are included in two or more frequencies of the same distribution. For example, many of the students in a group may report participation in more than one extra-curricular activity. If the total number of students reporting is used as the base, the per cents will total more than 100.¹ However, this is not a matter of concern if the fact of multiple participation is noted in presenting the data in a table. The statement that 75 per cent of the students participate in intramural athletics and 35 per cent participate in extra-mural athletics is not likely to be misleading. Such statements are applicable to other types of data where the response is multiple.

In combining per cents from several groups, the items should be weighted in terms of the sizes of their respective populations. The simple average of 15, 25, and 50 per cent is 30 per cent. Suppose, however, that the three populations are respectively 100, 500, and 1000. Then

$$\begin{array}{rcl} 100 \times .15 & = & 15 \\ 500 \times .25 & = & 125 \\ 1000 \times .50 & = & 500 \\ \hline 1600 & & 640 \end{array}$$

and 640 is 40 per cent of the cases in the total population.

When frequency distributions are expressed in different scale

¹ In such a case the sum of the per cents is without meaning and, hence, should not be given in the table.

units, comparison of their relative variabilities may be made by calculating coefficients of variability. The formula is

$$V = \frac{100\sigma}{M}$$

If one distribution has a mean of 50 and a standard deviation of 15, its coefficient of variability is 30. The relative variability of this distribution is the same as one having a mean of 5 and a standard deviation of 1.5. The coefficient of variability of both distributions is 30. It should be noted that in using this formula the assumption is made that the zero points on the scales of the compared distributions are true zero points. Hence, the use of the formula is justified to the extent that this assumption is satisfied. However, we make the same assumption in the use of the mean itself. If one has calculated the quartile points of a distribution, the following formula may be used to calculate a coefficient of variability.

$$V_Q = \frac{100(Q_3 - Q_1)}{Q_3 + Q_1}$$

If the distribution is symmetrical, $\frac{Q_3 - Q_1}{2}$ is equal to the median deviation of the distribution and $\frac{Q_3 + Q_1}{2}$ is equal to the median. When this condition is satisfied, the formula becomes:

$$V_Q = \frac{100MdD}{Md}$$

When the population from which data have been secured is not large, a frequency distribution is likely to exhibit marked irregularities that are not characteristic of a large population or universe. Hence, if the data collected are used as a basis for generalizing in regard to the shape of the distribution, it is desirable to estimate the probable shape for a very large population or universe. One method is to add each frequency to its adjacent ones and divide the resulting sums by three as in the

following illustration. The frequencies at each extreme are doubled, added to their neighboring frequencies, and the resulting sums divided by three.

	f	f (Smoothed)
50	1	2
45	4	3
40	4	3
35	1	4
30	7	5
25	7	6
20	4	5
15	4	4
10	4	3
5	1	2
0	1	1

The method is known as that of moving or rolling averages. Sometimes the frequencies are averaged in groups of five, seven, or nine. When a distribution has been treated in this way, it no longer represents the measures actually secured. High frequencies are decreased and low ones increased. If the method is used injudiciously, it may eliminate significant features. If used wisely, the result may be a distribution which better represents the true conditions in that some of the errors, or fluctuations due to chance or other extraneous causes, have been reduced. The method of rolling averages is much used in dealing with historical statistics, i.e., the statistics of trends. It is used in business statistics to eliminate seasonal or short-time cyclical variations from time series in order to reveal trends over longer periods of time.

Techniques for comparing frequency distributions. On pages 112-13 attention was called to the limitations of an average. Both the variability and any irregularities in the distribution are neglected. Hence, unless the distributions of the groups of data being compared are approximately equivalent in shape and variability, the use of only the central tendencies may result in an interpretation that is misleading or even erroneous. Even when the distributions are approximately equivalent in shape

and variability, a more meaningful interpretation will usually be secured by comparing the distributions. Two distributions may be compared by calculating the per cent of one that reaches or exceeds the central tendency of the other. Bi-serial r may also be used as a measure of the overlapping of two distributions. This statistic has been defined by the formula ¹

$$r_{\text{bis}} = \frac{M_2 - M_1}{\sigma} \frac{pq}{z}$$

in which M_1 and M_2 are the means of the two distributions, σ the standard deviation of the total distribution formed by combining the two given distributions, p is the per cent of the total distribution which is contributed by the distribution from which M_2 is calculated, $q = 1 - p$, and z is the height of the normal probability curve corresponding to the per cent p or q of the total area.²

¹ This formula is given by Kelley, T. L. *Statistical Methods*. New York: The Macmillan Company, 1923, p. 247.

For other formulae see McNamara, W. J., and Dunlap, J. W. "A Graphical Method for Computing the Standard Error of Bi-serial r ," *Journal of Experimental Education*, 2: 274-77, March, 1934.

² $p + q = 1.00$ in the above formula, not 100. The value of z may be obtained from an appropriate statistical table. If p is greater than .50 (if it is not, use q) subtract .5 from it and from this value find z in Table XII of Holzinger's *Statistical Tables for Students in Education and Psychology*. The Kelley-Wood table appearing as Appendix C, Kelley's *Statistical Method*, may also be used. Bi-serial r is also used to determine the correlation between success and non-success on a given test exercise and the criterion measures (see page 184) or, in general, to show the correlation between a trait measured quantitatively and another trait or characteristic which is classified in two qualitative categories. It is assumed that the distribution of the characteristic underlying the qualitative categories is normal and that the regression is linear. In the case of success and non-success on a test exercise, and in general, the means used in the above formula refer to the sub-series of quantitative measures classified under the two categories, σ refers to the total quantitative series, and p and q refer respectively to the proportions of cases under each category, the total frequency under each rubric being divided by N of the total quantitative series. For an illustration of the calculations, see Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, pp. 271-72.

Kelley, T. L. *Op. cit.*, pp. 245-49.

For a formula convenient to use in analysis of test items, a table, and a nomograph, see

Dunlap, J. W. "Note on Computations of Bi-serial Correlations in Item Evaluation" *Psychometrika*, 1: 51 f., June, 1936.

Symonds ¹ has compared measures of overlapping. For the case when the two distributions are normal and represent equal populations, he gives a table of the comparable values of (1) bi-serial r , (2) difference of the means of the distribution in terms of the standard deviation of one of them, and (3) the proportion of one distribution that is above the mean of the other. The use of bi-serial r is recommended.

Techniques involved in studying growth or trends. In studying the growth of the average achievement of a class, a survey is made at intervals over a period of months or years. The findings from such a series of surveys are frequently referred to as a time series. The interpretation is accomplished by comparison. If the average achievements of the class are represented as ordinates and the time intervals as abscissa distances, the curve joining the points thus located will represent the growth of the achievement of the group. The principal consideration in such studies is that the successive surveys be comparable. In studies of the trend of such phenomena as school enrollment, the data for surveys of past dates must necessarily be secured from records. In case there has not been uniformity in recording the data, the investigator faces the problem of estimating the adjustments necessary to make the measures comparable. Sometimes the growth in achievement or some other characteristic of children is studied by securing measures of a series of groups of children now in school. The possibility that these groups may not be comparable in certain significant respects makes it desirable to avoid this procedure when possible.

When the average rate of increase between two dates is desired, it cannot be obtained by dividing the total per cent of increase by the number of time intervals. It is necessary to calculate the geometric mean. Suppose for example that the enrollment of a given high school was five hundred in 1920

¹ Symonds, P. M. "A Comparison of Statistical Measures of Overlapping with Charts for Estimating the Value of Bi-serial r ," *Journal of Educational Psychology*, 21: 586-96, November, 1930.

and nine hundred in 1930. The total increase is 80 per cent of the initial enrollment. The average rate of increase for the ten-year period is *not* 8 per cent. Let r designate the average rate of increase and P_0 the enrollment in 1920. The enrollment in 1921, P_1 , will be equal to $P_0(1 + r)$. Similarly P_2 , the enrollment in 1922, will be equal to $P_1(1 + r)$ which by substitution becomes $P_0(1 + r)^2$. Continuing this reasoning we obtain $P_{10} = P_0(1 + r)^{10}$. This equation may be solved for r by employing logarithms.

$$\log P_{10} = \log P_0 + 10 \log (1 + r)$$

$$\log (1 + r) = \frac{\log P_{10} - \log P_0}{10}$$

Substituting the given values of P_0 and P_{10} and employing a table of logarithms, the value of r is found to be .0605 or 6.05 per cent, the average annual increase in the enrollment.

The mathematical relationship involved in the equation $P_{10} = P_0(1 + r)^{10}$ is that of compound interest. The final equation may be generalized by substituting n for 10.

$$\log (1 + r) = \frac{\log P_n - \log P_0}{n}$$

If the average rate of decrease is sought, the signs of the terms in the numerator of the fraction are changed.

School enrollments and census data over a period of years are frequently studied to determine some means for forecasting future enrollments. Several techniques¹ have been developed and employed in school surveys. Forecasting the future from the past is based upon the assumption that the trend of the past will be continued in the future. This is likely to be approximated, provided no new factors are introduced into the situa-

¹ For a brief description of the techniques, see

Engelhardt, Fred. "Forecasting School Population," *Teachers College, Columbia University Contributions to Education*, No. 171. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 66 pp.

Chamberlain, L. M., and Crawford, A. B. "The Prediction of Population and School Enrollment in the School Survey," *Bulletin of the Bureau of School Service*, Vol. 4, No. 3. Lexington, Kentucky: University of Kentucky, March, 1932. 27 pp.

tion. New laws relating to the school attendance, changes in the industrial activities of the community, and the like are likely to affect school enrollments and hence make the predictions less satisfactory than they would otherwise be. Chamberlain and Crawford¹ compared forecasts of school enrollments made in thirty-five city surveys with the actual enrollments in 1930-1931. Although the earliest of these forecasts was made in 1920 and the median date is 1924, the weighted mean error is 11.49 per cent. These authors recommend the use of simple and direct methods of forecasting.²

Graphic methods. The use of graphs to give meaning to summarizations of data is characteristic of reports of survey research, but a comprehensive treatment of the subject is unnecessary here since there are several excellent sources of information with respect to graphic methods.³ Frequency polygons, histograms, or column diagrams, smoothed frequency curves, and fitted ones are used in portraying frequency distributions. Line graphs are useful in indicating trends in enrollment, costs, expenditures, salaries, and the like, and in illustrating the development of traits.⁴ The typical curves used to

¹ *Op. cit.*, p. 22.

² The reader interested in further study of this topic should consult Chaddock, *op. cit.*, Chapter XIII or Smith, J. G. *Elementary Statistics*. New York: Henry Holt and Company, 1934, Chapters XI to XV. The treatment of the techniques in these sources is with reference to the field of economics in which the study of trends and forecasting are important topics.

³ See any standard text in statistics or educational statistics, or for more comprehensive treatments see

Alexander, Carter. *School Statistics and Publicity*. New York: Silver, Burdett and Company, 1919, Chapters IV and XI.

Brinton, W. C. *Graphic Methods for Presenting Facts*. New York: The Engineering Magazine Company, 1914. 371 pp.

Karsten, K. G. *Charts and Graphs*. New York: Prentice-Hall, Inc., 1923. 724 pp.

Williams, J. H. *Graphic Methods in Education*. Boston: Houghton Mifflin Company, 1924. 319 pp.

For an excellent discussion of the precautions to be observed in drawing graphs see Chaddock, *op. cit.*, Chapter XVI.

⁴ In some situations a logarithmic graph is helpful. See

Allen, C. B. "Logarithmic Charts," *Educational Administration and Supervision*, 20: 583-91, November, 1934.

Allen, C. B. "Rate of Change vs. Absolute Change in School Enrollments," *Educational Administration and Supervision*, 20: 431-37, September, 1934.

show how mental age increases with chronological age are examples of the latter. The familiar bar-graph is a modification of the histogram in which the columns are separated and are usually colored black. Frequently, the bars run horizontally from the vertical axis. The bars may represent frequencies in categories of an unordered series, and the order of the bars may be in terms of increasing or decreasing frequencies in the categories. Sometimes, bars in contrasting symbolism are drawn in pairs, or in threes. For example, if one wishes to present graphically data relative to the participation of boys and girls in extra-curricular activities, pairs of parallel black and white bars may be used for each of the activities represented in the data. The black and white bars of each pair may be drawn in contact with each other. The frequencies should be represented by the *lengths* and not by the areas of the bars. The bars should be in the same units, or according to the same scale,¹ and should be measured from the same zero point. If the lengths of the bars are not made proportional to the frequencies in this way, the graph will be misleading. The reader will be given the impression that differences in the lengths of the bars are more significant than they really are.

It is unwise to use circles of various sizes to indicate relative magnitudes. It is difficult to make accurate comparisons of such areas. Circles may be used, however, to illustrate the proportions of a population which fall into different categories. The area of the circle represents 100 per cent, and the sectors represent the per cents of each category. A circle divided in this way is known as a pie graph. Bars, or rectangles, may also be used to represent the per cents of a given sample which fall into different categories. When several circles, or bars, are used to represent several comparable percentage distributions, the dimensions of the bars or circles should be identical and the segments or sectors representing comparable per

¹ The nature of the scale should be indicated on the graph. A good procedure is to draw a line above or below the bars on which the scale values are indicated at intervals by short vertical lines accompanied by the values of the lower limits of the intervals.

cents should be given in the same order and in the same symbolism.

The ogive or percentile graph is used in portraying cumulative frequency percentages. Percentile ranks¹ may be read from percentile graphs and one can easily observe the per cents of cases above or below the median, the first and third quartiles, or the decile points. One can compare two percentile curves on the same chart with respect to the amount of overlapping of the distributions, their relative variability, and their respective central tendencies.²

Graphical representations of two or more distributions may be reduced to comparable scales and superimposed, but if the number of groups of data is large, the resulting figure will be complex and not easy to interpret. A simpler figure may be obtained by representing each distribution by a straight line on which the central tendency and certain percentile points are indicated. If these lines are arranged in parallel with respect to a common scale, the overlapping will be apparent. This type of representation is illustrated in Figure 3, on page 242, which is taken with minor changes from a Report by the Advisory Committee on College Testing in *The Educational Record* for October, 1932. In this figure the heavy portions of the vertical lines represent the range of the middle two-thirds of the distribution of the scores on the English test administered to sophomores in the colleges represented. The narrow extensions show the range from the tenth percentile to the ninetieth percentile. The average score for each college is represented by a short horizontal line.

¹ The percentile rank of a measure is the per cent of measures below that measure in the distribution. The percentile point nearest which the measure falls may be taken as its percentile rank, but if precise values are needed, they should be calculated by means of a formula. See Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, pp. 136 f.

² For discussion of the methods employed in drawing percentile graphs, see Holzinger, *op. cit.*, pp. 127-40.

Odell, C. W. *Educational Measurement in High School*. New York: The Century Company, 1930, pp. 610-15.

Otis, A. S. *Statistical Method in Educational Measurement*. Yonkers-on-Hudson, New York: World Book Company, 1925, pp. 53-67, 77-84, 95-100.

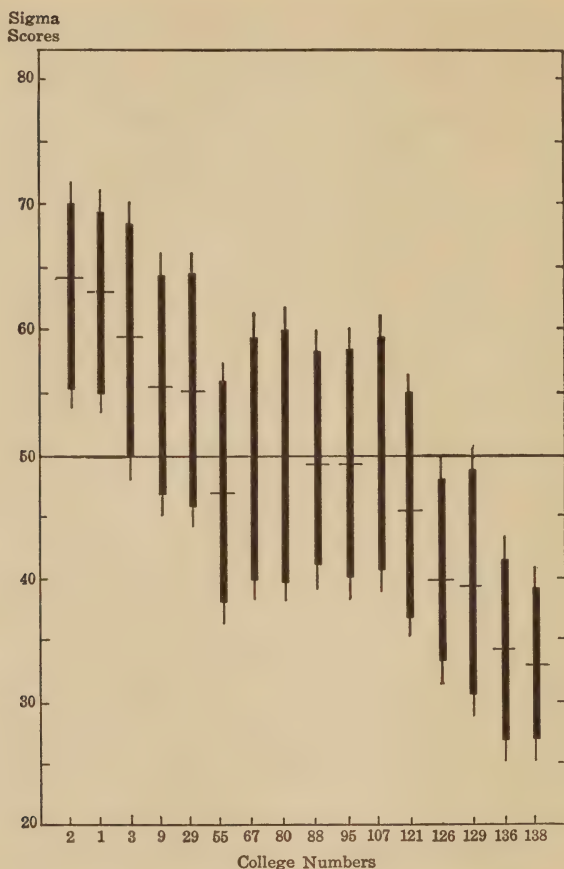


FIG. 3. Bar graph of variability of English achievement in several colleges. The "college numbers" designate 16 of the 138 colleges participating in the testing program. After Johnston, J. B., *et al.*, "The 1932 College Sophomore Testing Program," *Educational Record*, 13:306, October, 1932.

B. REPORTING AND INTERPRETING FINDINGS

Reporting survey studies. In reporting a survey study an investigator faces the problem of deciding how much of the

details to include. In dealing with this problem the principal considerations are: (1) the availability of space, (2) the probable interest of readers in details, and (3) the effectiveness of the report in fulfilling its communicative function. If the survey is being reported in unpublished form, space is seldom an important consideration; but when the report is to be published, it is usually desirable to limit frequency tables and other details to a minimum. The critical reader is likely to be interested in details as a means of determining the dependability of the survey, but most readers are not interested in details. They wish to learn the principal findings and their interpretation. Hence, the author should be guided by the audience for which he is writing. Comprehensive surveys covering large areas are more likely to command the attention of critical readers than minor studies. The presence of many detailed tables tend to interfere with a fluent reading of the report, but frequently a survey cannot be effectively reported without presenting a number of details. For example, if per cents have been calculated from the frequencies of a distribution, it is usually desirable to report the absolute frequencies as well as the relative ones.

Few definite rules can be stated. The last sentence of the preceding paragraph is one generally recognized. Another rule is that the details of calculation should not be reported, but, if the statistical procedure employed is not a standard one, it should be described. Usually a measure of the variability of a frequency distribution should be given as well as its central tendency. This is important when precise comparisons are being made. If sampling has been employed in collecting data or it is desired to generalize from the findings, information relative to the representativeness of the data should be given. The calculation of the probable error of a mean or median is justified only when a process of random sampling has been employed in collecting the data or when the group of data may be assumed to constitute a random sample. Hence, this statistic should not be given unless one or the other of these conditions

exist. Unfortunately the probable error is frequently introduced in survey investigations when no sampling has occurred or when the sample is obviously not random. For example, in a study ¹ in which no sampling occurred, the following statements are made. "The arithmetic mean of the grades earned in residence is 2.33 ± 0.043 . The mean of the grades earned by correspondence is 1.95 ± 0.032 ." Interpreted literally, these statements are absurd.

If it seems desirable to include a large number of details, tables that are likely to be of interest only to the more critical readers may be placed in an appendix. The reading of a simple table need not be described in the accompanying text, but when the meaning of the entries will not be obvious to a competent reader, their interpretation should be given. In all cases the captions of the table, especially those of the columns, should be formulated with care.

Interpretation of survey findings. A statement of the median score made on a certain achievement test by the seventh-grade pupils in City A is of little interest. The statement that the median score in City A is ten points less than in City B is more meaningful, but usually it is desired to say that the difference is an index (indirect measure) of the relative achievement of the pupils or of the relative quality of the instruction they have received. Hence, the interpretation of the findings in an achievement survey may be thought of as involving a comparison and usually a change in the label attached to the difference. The procedure of the interpretation in other types of surveys is similar.

Dependability of interpretations when generalization is not involved. When a difference is labeled as an index (indirect measure) of a trait or condition, it is necessary to justify this interpretation. In Chapter V, the term *dependability* was introduced to refer to the degree of correctness of a statistic when the

¹ Larson, E. L. "The Comparative Quality of Work Done by Students in Residence and Correspondence Work," *Journal of Educational Research*, 25: 105-09, February, 1932.

precise nature of its label is considered. Hence, the justification of an interpretation may be thought of as demonstrating the dependability of a difference when it is given the desired label. This may be accomplished by showing that the difference cannot be satisfactorily explained as the sum of contributions from other possible causes. For example, justification of the interpretation of a difference between mean test scores as an index of relative achievement is accomplished by showing that it is very improbable that the difference is due to the effects of data faults. If the interpretation is in terms of relative quality of instruction, it is necessary to consider also the quality of the pupil material and the amount of instruction to which they have been subjected. If the interpretation is generalized, attention must be given to the representativeness of the groups tested.

The details of demonstrating the dependability of an interpretation vary with the nature of the data and the desired interpretation, but the general procedure may be illustrated by considering the interpretation of differences between mean test scores as measures of relative achievement. The effect of variable errors upon a mean varies inversely with the square root of the number of cases. Hence, the contributions from variable errors of measurement and variable errors of validity will be small, and if the number of cases is large, they may usually be considered negligible. This means that the unreliability of the test and its lack of validity, as the term is usually defined, are relatively unimportant in surveys and consideration of them may be omitted without seriously weakening the argument. Frequently an investigator points with pride to the coefficient of reliability of the test used and thereby gives the impression that consequently his interpretation is highly dependable. This is unfortunate because variable errors of measurement whose magnitude is indicated by a coefficient of reliability usually make only a minor contribution to a difference between means or medians. Similar statements may be made with reference to indices of validity.

The contributions from systematic errors of measurement may

be large. As pointed out in Chapter V, there is no definite procedure by means of which the magnitude of the systematic error of measurement in a group of test scores may be determined. We may, however, obtain some indication of their magnitude by inquiring concerning the testing conditions which include the explanation of the test to the pupils, the attitude of the person administering the test, the number of minutes allowed for responding to the exercises, and other aspects of the administration of the test that influence the mean score of a group. If it does not appear that the testing conditions were the same for the populations tested, it is likely that a portion of the difference is due to a systematic error of measurement.

When the label attached to the difference or implied in its interpretation specifies an achievement other than that measured directly by the test, the possibility of a systematic error of validity is created. It is seldom that the achievement directly measured by a test is the achievement whose specification is desired in the interpretation. For example, a sentence dictation spelling test measures directly the ability of the pupils to spell certain words under the conditions of the test. We usually desire, however, measures of the ability to spell the words used in typical writing. Hence, when scores from a dictation spelling test are labeled "measures of spelling ability" the possibility of errors of validity is created. The ability to respond to a silent reading test is not the same as the ability that functions in typical reading. The ability to identify statements that are true and those that are false is not what we usually mean by achievement in a school subject. In addition to the substitutions implied by these statements, we frequently substitute measures of a sample of achievement for measures of the total achievement in a school subject or in a specified segment of it. The tests that we designate as measuring achievement actually measure a combination of achievement and general intelligence.

In a given situation the mean status of what the test measures directly bears a certain ratio to the mean status of the achievement whose measurement is desired. If this ratio is the same

for two populations, the change of label made by the interpretation will not introduce a systematic error. For example, under a given curriculum and plan of instruction in arithmetic, the average calculation achievement of a group of pupils bears a certain ratio to their average total achievement in arithmetic. If the same balance of curriculum and instruction prevails in two populations, the difference between their mean scores on a calculation test may be labeled "difference in arithmetical achievement" without introducing a systematic error of validity. If, however, the curriculum and general plan of instruction are different in the two populations, a systematic error of validity is likely to be introduced when the difference between mean scores on a calculation test is labeled "difference in arithmetical achievement."

If the difference between mean test scores is interpreted as an index of the relative quality of instruction, it is necessary to consider also the equivalence of the two groups with reference to capacity to learn and the amount of instruction received. Measures of capacity to learn are provided by intelligence test scores, but they may involve a systematic error. Hence, in demonstrating the equivalence of the groups compared by introducing scores on an intelligence test, it is necessary to inquire concerning the possibility of a systematic error in these measures. In determining the amount of instruction, it is necessary to consider incidental instruction as well as that for which explicit provision is made.

Comparisons with norms are especially difficult to interpret because so little information is provided concerning the testing conditions and the populations from which the norms have been derived. A grade population is indefinite, especially at the high school level. Assuming that the population from which the norm was derived is typical of the grade with respect to general intelligence and previous training in the general field of the test, it may not be typical with reference to the acquaintance of the pupils with testing procedures in general or with respect to tests of the type administered. Hence, the dependability of

interpretations involving comparisons with norms is usually uncertain. One writer has presented evidence to show that "all but a very few of the norms now available for high school tests" are "practically worthless for the evaluation of school achievement on a relative basis."¹

Dependability of generalized interpretations. Frequently, only a sample of the population specified by the problem is measured or it is desired to generalize from the findings of a survey. If the sample is perfectly representative, the dependability of the interpretation will not be affected when the findings are labeled as measures of a larger population or universe. If the sample is not representative, this condition may contribute an additional error to the findings as measures of the larger population or universe.

Sometimes it is possible to present evidence to show that a sample is highly representative. For example, in his study of the ability of ninth-grade pupils to read typical literary selections, Irion² was able to show that the group tested was typical of ninth-grade pupils in general. Hence, generalization of his interpretations did not materially affect their dependability.

In many cases, the population surveyed is obviously not representative or the method of collecting the data is such that non-representativeness is highly probable. When the per cent of returns in a questionnaire study does not approximate one hundred, the data are frequently not representative. What is likely to happen is illustrated by a study in which a questionnaire was mailed to the graduates of the College of Education at the University of Illinois for the purpose of ascertaining the proportion of the graduates who enter and continue in educational

¹ Lindquist, E. F. "Factors Determining Reliability of Test Norms," *Journal of Educational Psychology*, 21: 512-20, October, 1930.

Another pertinent reference on this point is Stokes, C. N., and Finch, F. H. "A Comparison of Norms on Certain Standardized Tests in Arithmetic," *Elementary School Journal*, 32: 785-87, June, 1932.

² Irion, T. W. H. "Comprehension Difficulties of Ninth Grade Students in the Study of Literature," *Teachers College, Columbia University Contributions to Education*, No. 189. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 116 pp.

work. Returns were received from approximately 53 per cent of those graduating during the period from 1920 to 1930. It seemed reasonable that graduates engaged in educational work would be more likely to respond to the questionnaire than those not so employed. This hypothesis was supported by data obtained from the *Alumni Directory*. Hence, the per cent of those reporting who were engaged in educational work was greater than the proportion of all graduates engaged in educational work. When the nature of the selection is known, as in this case, it may be possible to make estimates that will be more dependable than the results obtained from the data.

Occasionally, it is possible to employ a technique of random sampling in collecting survey data and in certain situations the assumption of a random sample may be justified. In such cases probable or standard error formulae may be employed as a means of calculating the probable effect of chance non-representativeness.¹ It is a common practice to compare a difference with its probable error as a means of determining its statistical significance. It is called statistically significant if the probability of the effect of chance non-representativeness being large enough to change the sign of the difference is very small, usually not more than 3 or 4 out of 1000.² The statistical significance of a difference has nothing to do with the contributions from systematic errors or other causes. Hence, a statistically significant difference is not necessarily dependable. It is unfortunate that the use of the probable error and its derivative the critical ratio is frequently discussed under the head of

¹ See pages 104-05 for formulae.

² This probability is conveniently obtained by means of the critical ratio (*CR*) which is formed by dividing the difference by its probable error. Tables have been prepared which give for various values of the critical ratio the probabilities that the difference for the universe will have the same sign. For example, see Garrett, H. E. *Statistics in Psychology and Education*. New York: Longmans, Green and Company, 1926, p. 135.

The term *critical ratio* was first used in McGaughy, J. R. "The Fiscal Administration of City School Systems." New York: The Macmillan Company, 1924, p. 9. The concept of the critical ratio, however, had been employed by mathematicians for some time. The experimental coefficient (see Chapter IX, page 308), proposed by McCall, is based on the same principles.

reliability, thereby giving the impression that it is only necessary for an investigator to consider the probable error when seeking to determine the dependability of survey findings.

It is seldom feasible to employ a technique of random sampling in educational research.¹ Hence, usually it will be necessary to justify the use of probable or standard error formulae by showing that the groups of data may be regarded as random samples. The formulae should not be used when the sample can be shown to be highly representative or when the evidence indicates that it is non-representative. When a survey is made within a school system or within a limited area, it is seldom justifiable to assume that the data collected constitute random samples of school systems in general. If the high school principals of a state are invited to administer a general achievement test to their twelfth-grade students, it is likely that a large proportion of the favorable responses will be received from the better schools. Hence, it should not be assumed that the group of schools administering the test form a random sample of the state. On the other hand, it may be argued that the twelfth-grade group within a given high school may be considered a random sample of the twelfth-grade students in this school over a period of years, provided no systematic influence is apparent. If the rules relative to age at entrance to the first grade, the promotion policy, and other conditions affecting the age of high school seniors have not varied, the group may be considered a random sample relative to chronological age. It should not be considered a random sample relative to school achievement if significant changes have been made in the curriculum or other conditions intended to increase the achievement. Although the assumption that a sample is random may appear to be justified, unobserved factors may make the sample a non-random one. Chaddock² has pointed out that the application of probable error techniques is probably not justified in the

¹ For an elaboration of this point, see pages 58-59 and 109.

² Chaddock, R. E. "Significance of Infant Mortality Rates for Small Geographic Areas," *Journal of American Statistical Association*, 29: 243-49, September, 1934.

case of infant death rates for adjacent areas. In other words, the numbers of infants dying during a given year do not appear to be random samples of the number of infant deaths over a period of years. Chaddock describes a case in which determination of statistical significance resulted in an erroneous interpretation.

Further consideration of the dependability of interpretations.

Interpretation may be thought of as an attempt to explain survey findings in terms of their causes. As an illustration, consider a statewide or regional survey in which a test or battery of tests is administered to the twelfth-grade pupils in the high schools of the area. The mean scores of the schools will usually exhibit a wide range of variability such as is shown in Figure 3 on page 242. Lindquist ¹ has reported that in a statewide survey of achievement in Iowa high schools, the variability of the mean scores was approximately half of the variability of the individual scores. Such findings suggest similar differences in mean achievement and in school efficiency, but this hypothesis must be tested by estimating how much of the variability of the mean scores can be explained as being due to other causes. If it appears at all likely that the variability is due to other causes, the interpretation that there are differences in mean achievement or in school efficiency is not justifiable. If it appears that a large proportion of the variability is due to other causes, then it must be concluded that the differences in mean achievement or in school efficiency are much less than the findings indicate. If the magnitude of the contributions from other causes is uncertain, then the interpretation must be considered uncertain.

If the interpretation is generalized, that is, if the calculated statistics are considered typical of the schools over a period of years, the non-representativeness of some of the groups of pupils tested is a possible cause of the variability of the means. If careful inquiry reveals no information to the contrary, it may be assumed that the twelfth-grade pupils of a given year constitute

¹ Lindquist, E. F. "Factors Determining Reliability of Test Norms," *Journal of Educational Psychology*, 21: 512-20, October, 1930.

a random sample of the twelfth-grade pupils of a school over a period of years. On the basis of this assumption, a portion of the variability of the mean scores may be explained as being due to the effect of chance. In his study, Lindquist estimates that chance might make the variability of the mean scores about one-fifth as great as that of the individual scores. In the case of some of the pairs of schools, chance is likely to be responsible for a large portion of the observed difference.

If the range of the mean scores is relatively small, it may be desired to ascertain whether their variability may be explained as probably due to chance alone. One method is to calculate the correlation ratio which may be obtained from the following relationship.¹

$$\eta^2 = \frac{\sum N_i(M_i - M_t)^2}{\sum (X_i - M_t)^2}$$

The numerator of the fraction is the weighted sum of the squares of the deviations of the means of the several groups from the mean of the total distribution (M_t) and the denominator is N_t times the square of the standard deviation of the total ($N_t\sigma_t^2$). After η has been calculated, it may be tested for statistical significance.² If the correlation ratio is not significant, it may be concluded that the variability of the means is explainable as possibly due to chance. Fisher³ has developed a somewhat

¹ This formula can be derived from that given on page 98. See also page 88.

² For an illustration of this method, see Reitz, Wilhelm. "Statistical Techniques for the Study of Institutional Differences," *Journal of Experimental Education*, 3: 11-24, September, 1934.

³ Fisher, R. A. *Statistical Methods for Research Workers*. London: Oliver and Boyd, 1928, Chapter VII.

See also Snedecor, G. W. *Calculation and Interpretation of Variance and Covariance*. Ames, Iowa: Collegiate Press, Inc., 1934, Part I.

For applications of Fisher's technique, see

Reitz, Wilhelm, *op. cit.*

Lyon, V. E. "The Variation of High School Senior and College Freshman Classes," *Journal of Experimental Education*, 3: 25-35, September, 1934.

Lyon gives mean intelligence percentile scores for 108 high schools for five successive years. The means for a given school vary from year to year. In some cases, the range is relatively large. Chance is doubtless the principal cause of this variation, but it may be contributed to by systematic influences. The reader who consults these references should note that neither of the authors appears to recognize the possibility that the variability of the means of a group of schools

different technique. It should be noted that demonstration of the statistical significance of η does not prove the dependability of the interpretation. When there is reason to expect that the means are subject to systematic errors of either measurement or validity or both, a crude estimate of the probable effect of chance, such as Lindquist made, will be a more helpful procedure.

In a coöperative testing program there are likely to be some variations in testing conditions which will contribute to the variability of the mean scores. Even when there is a conscientious attempt to follow the letter of the instructions, there may be accidental variations and differences in the attitude of the examiners which will result in systematic errors of measurement. Furthermore, since the measurement of achievement is indirect, errors of validity are possible and variations in the curriculum and general plan of instruction ¹ are likely to introduce systematic errors of validity when the scores are labeled measures of achievement. If the interpretation is in terms of relative efficiency of the schools, the total effect of the systematic errors of measurement and the systematic errors of validity may be sufficient enough to explain many of the larger differences between mean scores. Hence, a large proportion of the variability of the mean scores which may appear as very astonishing to an uninformed person is probably due to systematic errors of measurement, systematic errors of validity, and non-representativeness. However, the uncertainty in regard to the magnitude and direction of the contributions from these sources makes the interpretation of the findings of such a survey very hazardous. It may be that if corrections could be made for the effect of non-representativeness, systematic errors of measurement, and systematic errors of validity, a school occupying a relatively low

for a single test may be contributed to by errors that are systematic within schools.

¹ The objectives towards which instruction is directed may vary not only from school to school but also within the same school from year to year.

Remmers, H. H. "The Stability of Relative Excellence of High Schools," *School and Society*, 38: 412-16, September 23, 1933.

position in the distribution of mean scores would be shown to belong near the top of the distribution.¹

As another illustration, consider a survey of the duties performed by high school principals. A request to keep a diary of their duties for a period of two weeks is sent to two hundred high school principals selected at random from a list of the accredited schools within a certain area. Apparently complete diaries are received from fifty-seven principals. From this information, there is compiled a list of the duties reported and the average number of minutes per week devoted to them by the fifty-seven principals. In interpreting these findings as "distribution of time of typical high school principal" or as "indices of the importance of duties performed by high school principals," several possibilities of error must be considered. It is not likely that the records kept by the principals are subject to a systematic error in the case of the more overt duties. The records, however, may be in error in the case of duties that involve considerable deliberation such as the planning of school policies. Unless the terms employed to designate the duties are carefully defined, it is likely that there may be some misunderstanding in regard to what the designated duty included. Furthermore, the duties performed during the two weeks may not be representative of the school year and the group of principals reporting data may not be typical. Non-representativeness in either case will affect the dependability of generalizations. Hence, the dependability of an interpretation of the findings as "distribution of time of a typical high school principal" should be regarded as uncertain. An interpretation of the average number of minutes per week as indices of the relative importance of the duties is likely to be less dependable.

A determination of the consensus of opinion relative to certain practices or issues is frequently attempted by employing a ques-

¹ The uncertainty of the dependability of the interpretation of the findings in statewide or regional surveys of school achievement constitute a strong argument against such surveys. There are, however, other considerations. See Douglass, H. R. "The Effects of State and National Testing on the Secondary School," *School Review*, 42: 497-509, September, 1934.

tionnaire consisting of groups of statements each of which expresses a number of opinions relative to a given topic or question. The recipient is asked to check within each group the statement that most nearly represents his opinion or belief relative to the designated topic or question. Suppose such a questionnaire is mailed to one hundred "representative" educators and that usable replies are received from forty. In considering the dependability of the statement checked most frequently as the consensus of opinion of "representative" educators, attention should be given to the representativeness of the population responding to the questionnaire and to the accuracy of the statements checked as expressing the opinions of the several correspondents.

The fact that a person occupies a position of prominence or has attained a reputation for good judgment does not make him competent with reference to all questions and issues. It is a frequent observation that a person whose acquaintance with a field is casual is more willing to express an opinion and is more certain in his beliefs than an intensive student of the field. Furthermore, if several persons, regarded as authorities in a given field, are asked for an opinion with reference to a particular question or issue, some of them will probably give the matter careful consideration while others will answer in a casual manner. In such a case, the authorities responding should not be regarded as equally competent. In other words, the competence of a person depends in part upon the manner in which he responds. A person should not be regarded as competent merely on the basis of his general reputation or his reputation in a field not identified with the particular question or issue being studied. Hence, it is likely that the forty persons replying to the request are not representative of "competent" educators in general.

When the most frequently checked statements are designated as the most correct beliefs or opinions, there is the implication that the judgments in one direction from the truth balance those in the opposite direction. This is likely to be the case when a physical magnitude such as the distance between two points is

estimated by a number of competent persons and the average of the estimates is likely to approximate the true magnitude. This balance of deviations from the truth may not prevail when opinions or beliefs are expressed relative to educational questions, especially when controversial matters are involved. Tradition, prevailing public opinion, or other influences may cause the expressed opinions to be biased, even when statements have been secured from persons regarded as competent. For example, expressions concerning the value of Latin as a secondary school subject secured from teachers of Latin, especially those occupying positions of prominence, will be biased. A consensus of opinion relative to the desirability of German as a secondary school subject obtained during the academic year of 1917-1918 would not have been dependable. In other words, opinions may involve a systematic error and when this occurs the consensus of the opinions expressed should not be regarded as the truth or the wisest belief.¹

If a series of surveys are being utilized as a basis for studying trends, the apparent changes may be misleading. For example, consider a study in which a test is administered to pupils in grades six to twelve inclusive as a means of ascertaining the growth in language ability. In interpreting the differences between the mean scores of the successive grades as indices of the growth in language ability, it is necessary to consider the probable contributions from other sources. Chance may make some of the differences larger, but these will likely be balanced by ones made smaller. Hence, chance will not be likely to affect the general trend. Neither will the general trend be likely to be affected by systematic errors of measurement. Due to the curriculum, the general plan of instruction, and the increasing maturity of the pupils, the ratio of what the test measures directly to language ability in grade six may not be the same as the corresponding ratio for the years of the senior high school. Hence, the trend defined by the mean scores for the successive

¹ For a discussion of consensus of opinion in curriculum construction, see Chapter XII.

grades may not correctly represent the growth in achievement. There may be additional error due to the selection of the successive grade groups. If the trend in the enrollment of a school is being studied, attention must be given to the comparability of the data for the successive years.

Appraisal of survey findings in terms specifying degrees of value. Comparison of survey findings from two or more populations or units, or comparison with norms may be thought of as appraisal in terms of more or less. Frequently an appraisal in terms of value is desired. Studies reported by Fowlkes ¹ and Heck ² afford illustrations of this type of appraisal. In both researches, interpretations of the data collected are made in terms of what practices are to be held superior.

Appraisal of this type requires a criterion, i.e., specification of what is desirable or what should be. Such criteria cannot be determined objectively. An investigator may survey and determine the "average" present practice or ascertain a consensus of opinion of authorities, but somewhere in the process he must decide what the criterion is to be. The report of Thomas ³ on public school plumbing equipment is an excellent illustration of the derivation of criteria. Thomas studied the literature dealing with plumbing equipment for schools, interviewed experts, and observed the plumbing in schools, office buildings, hotels, and public comfort stations. With the information from these sources at hand, he formulated the criteria which he used in the evaluation of present practices in public school plumbing equipment and in making recommendations relative to the plumbing equipment of new school buildings.

Accounting systems and plumbing equipment are tangible

¹ Fowlkes, J. G. "The Accounting of Public School Expenditures in Wisconsin," *University of Wisconsin, Bureau of Educational Research Bulletin*, No. 4. Madison: University of Wisconsin, 1924. 59 pp.

² Heck, A. O. "A Study of Child-Accounting Records," *The Ohio State University Studies*, Vol. 2, No. 9, Bureau of Educational Research Monographs, No. 2. Columbus: Ohio State University Press, 1925. 245 pp.

³ Thomas, M. W. "Public School Plumbing Equipment," *Teachers College, Columbia University Contributions to Education*, No. 282. New York: Bureau of Publications, Teachers College, Columbia University, 1928. 128 pp.

and the criteria arrived at by a discerning and critical student of present practices are likely to be acceptable to most persons. Acceptable criteria for appraising pupil achievement are more difficult to determine. Lists of educational objectives are available, but authorities do not agree, especially when the objectives are sufficiently detailed to serve as a basis for evaluating pupil achievement. Test norms indicate the achievement that characterizes the average or typical pupil, but they fail to indicate whether or not greater achievement is to be valued. High achievement in handwriting may be decreasingly desirable in a civilization which values less and less the ability to write attractively as well as legibly. High achievement in Greek or Latin may be decreasingly desirable in a civilization which places a decreasing value on a knowledge of these languages.

It is somewhat futile to attempt to estimate the possible dependability of appraisals of present practices and conditions in terms of value or desirability. They may be regarded as "sound" or "justified" to the extent that it is evident that they are made with due consideration of all relevant factors. A further test of the justification of the appraisals of this type is the extent to which they are instrumental in the improvement of practices or conditions, but in many cases it would be difficult to determine what constitutes improvement. The possibilities and limitations of methods that may be employed in the solution of problems implying a determination of "what should be" are considered at greater length in Chapter XII.

The identification of relationships. The study of trends referred to on page 215 may be thought of as an attempt to determine the relation between time and the status being studied. Other types of survey data are frequently studied to determine relationships. As pointed out on page 215, such studies are beyond the scope of this chapter when correlation analysis is employed, but it may be noted that in several cases the interpretation of comparative survey findings implies the recognition of a relationship and an explanation of a difference

in status is essentially an attempt to identify one or more causal relationships.

ILLUSTRATIVE BIBLIOGRAPHY OF SURVEY INVESTIGATIONS

This illustrative bibliography is supplementary to surveys mentioned in the preceding pages. In compiling it the authors have endeavored to include illustrations of a variety of techniques for collecting data rather than to present a list of model surveys. A number of the studies are subject to criticism, in some cases rather serious ones. The annotations are intended to indicate the general procedure of the survey. No attempt is made to describe the findings or to evaluate them. A number of survey studies relating to the curriculum are given in the bibliography of Chapter XII.

1. ANDERSON, E. M. "Individual Differences in the Reading Ability of College Students," *University of Missouri Bulletin*, Vol. 29, No. 39, Education Series No. 25. Columbia, Missouri: University of Missouri, 1928. 77 pp.

Data were collected by administration of tests. The representativeness of the 237 college students is considered.

2. ANDRUS, RUTH. "A Tentative Inventory of the Habits of Children from Two to Four Years of Age," *Teachers College, Columbia University Contributions to Education*, No. 160. New York: Bureau of Publications, Teachers College, Columbia University, 1924. 50 pp.

Data were secured by systematic observation of nursery school children and through examination of unpublished data and descriptions of observed behavior recorded in over one hundred books and articles.

3. BAIRD, D. O. "A Study of Biology Notebook Work in New York State," *Teachers College, Columbia University Contributions to Education*, No. 400. New York: Bureau of Publications, Teachers College, Columbia University, 1929. 118 pp.

Fifty-two representative biology notebooks were analyzed with respect to nature and number of experiments reported, kind and amount of drawing, vocabulary used, quality of handwriting, and certain other items. A basis of appraisal was effected by securing data from several sources.

4. BARR, A. S. *Characteristic Differences in the Teaching Performance of Good and Poor Teachers of the Social Studies*. Bloomington, Illinois: Public School Publishing Company, 1929. 127 pp.

Data were collected by general observations, attention chart, time chart, stenographic report, and by letters from superintendents and teachers.

5. BARR, A. S., and GIFFORD, C. W. "The Vocabulary of American

History," *Journal of Educational Research*, 20: 103-21, September, 1929.

Eight representative American history texts were subjected to vocabulary analysis, without sampling. Certain words were excluded on the basis of ten criteria, one of which was the first three thousand words of the Thorndike Wordbook list.

6. BARTLETT, L. W. "State Control of Private Incorporated Institutions of Higher Education," *Teachers College, Columbia University Contributions to Education*, No. 207. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 95 pp.

The sources of data consulted in this study include decisions of the United States Supreme Court, laws of states governing the incorporations of institutions of higher education, and charters of selected private colleges and universities.

7. BENDER, J. F. "The Functions of Courts in Enforcing School Attendance Laws," *Teachers College, Columbia University Contributions to Education*, No. 262. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 187 pp.

Sources of data include reports of state and city superintendents, state school surveys, court records, and state school laws. The author attended hearings of more than two hundred cases involving violation of attendance laws.

8. BENNETT, H. E. "A Study of School Posture and Seating," *Elementary School Journal*, 26: 50-57, September, 1925.

Facts and opinions concerning seating were collected from the literature on school hygiene. Medical literature was consulted with respect to information on the relations of sedentary postures and physical defects. Extensive and intensive observational studies were made of school children in which five thousand individual posture records were obtained. Measurements of more than 2500 children in all proportions pertinent to seating and desk dimensions were made by means of an elaborate apparatus devised for the purpose.

9. BLANKENSHIP, A. S. "The Accessibility of Rural Schoolhouses in Texas," *Teachers College, Columbia University Contributions to Education*, No. 229. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 62 pp.

The investigator compiled maps from school census data and data in the offices of county superintendents showing locations of rural school sites and rural dwellings. A questionnaire was sent to each of the schools of Texas employing transportation. Facts pertaining to area, length of transportation routes, and number of teachers employed were requested.

10. BOOK, W. F., and HARTER, R. S. "Mistakes Which Pupils Make in Spelling," *Journal of Educational Research*, 19: 106-18, February, 1929.

The sources of data used in this investigation were 3096 spelling test papers of pupils in the second to eighth grades, 608 compositions of first and second year high school pupils, and 1492 themes of college freshmen.

11. BUCKNER, M. A. "A Study of Pupil Elimination in the New Haven High School," *School Review*, 39: 532-41, September, 1931.

In addition to obtaining data from school records, the investigator employed the interview technique with respect to a sample of 196 pupils that had dropped out of school.

12. CALIVER, AMBROSE. "A Personnel Study of Negro College Students," *Teachers College, Columbia University Contributions to Education*, No. 484. New York: Bureau of Publications, Teachers College, Columbia University, 1931. 146 pp.

Data were collected by means of questionnaires, by examination of records, and by administration of tests.

13. CHADSEY, C. E., et al. "The Status of the Superintendent," *First Yearbook of the Department of Superintendence*. Washington: National Education Association, 1923. 206 pp. For description see page 2.
14. CHAPMAN, J. C., and EBY, H. L. "A Comparative Study, by Educational Measurements, of One-Room Rural-School Children and City-School Children," *Journal of Educational Research*, 2: 636-46, October, 1920.

In the interpretation of the data, interesting use is made of the Pearson Coefficient of Variation $\frac{100Q}{M}$. (In the formula ordinarily used, Q is replaced by σ .)

15. CHRISTOFFERSON, H. C. "College Freshmen and Problem Solving in Arithmetic," *Journal of Educational Research*, 21: 15-20, January, 1930.

A standardized arithmetic test was administered to a group of 99 college freshmen in this study. The investigator reports an analysis of the errors made on the test and also a comparison of the 25th percentile, median, and 75th percentile of those students with the corresponding eighth-grade norms.

16. COUNTS, G. S. "The Social Composition of Boards of Education," *Supplementary Educational Monographs*, No. 33. Chicago: University of Chicago Press, 1927. 100 pp.

Questionnaires were sent to superintendents to secure factual data relative to their boards of education and returns were secured from 1654 superintendents widely distributed about the country.

17. ELDER, VERA, and CARPENTER, H. S. "Reading Interests of High School Children," *Journal of Educational Research*, 19: 276-82, April, 1929.

A questionnaire was used in this study to secure data relative to the reading interests of 487 girls on all grade levels of one city high school. These data were supplemented by information obtained from reports of outside reading done by these students.

18. ELSBREE, W. S. "Teacher Turnover in the Cities and Villages of New York State," *Teachers College, Columbia University Contributions to Education*, No. 300. New York: Bureau of Publications, Teachers College, Columbia University, 1928. 88 pp.

Data were collected by means of a report blank filled out by 71 of the 83 superintendents in New York State, exclusive of New York City and Buffalo. Additional data were secured from records in the Teachers Retirement Bureau and State Department of Education.

19. FITCH, H. N. "An Analysis of the Supervisory Activities and Techniques of the Elementary School Training Supervisor in State Normal Schools and Teachers Colleges," *Teachers College, Columbia University Contributions to Education*, No. 476. New York: Bureau of Publications, Teachers College, Columbia University, 1931. 130 pp.

A check list of supervisory activities was prepared by examination of student teaching manuals and of previous research reported in the field. The list of 422 items was distributed to 779 supervisors. Usable returns were received from 355 supervisors.

20. FLOWERS, I. V. "The Duties of the Elementary-School Principal," *Elementary School Journal*, 27: 414-22, February, 1927.

The investigator sent a request to 170 principals of elementary schools to keep diaries of their daily work for a period of two weeks. Replies were received from 67 principals.

21. FOSTER, J. C. "Distribution of the Teachers' Time among Children in the Nursery School and Kindergarten," *Journal of Educational Research*, 22: 172-83, October, 1930.

The data were collected by observation in which the record blank included the following items: Name of child, activity of child, whether child came to teacher of his own accord, and the number of seconds spent by the teacher on the child.

22. HENRY, N. B. "A Study of Public School Costs in Illinois Cities," *The*

Educational Finance Inquiry, Vol. 12. New York: The Macmillan Company, 1924. 82 pp.

Data were collected by examination of the records of twelve Illinois cities ranging in size from eleven to seventy-six thousand. Supplementary information was secured through interviews. The study is interesting with respect to care used in reducing data to comparable bases.

23. HEWITT, ALDEN. "A Comparable Study of White and Colored Pupils in a Southern School System," *Elementary School Journal*, 31: 111-19, October, 1930.

The Illinois Examination, with the exception of its arithmetic test, was administered to ninety colored and to eighty-five white, seventh-grade pupils.

24. INMAN, J. H. "The Training of Iowa High School Teachers in Relation to the Subjects They Teach," *University of Iowa Studies in Education*, Vol. 4, No. 9. Iowa City, Iowa: University of Iowa, 1928. 66 pp.

Questionnaires were sent to 2000 graduates of eleven colleges in Iowa who had had from one to five years of high school teaching experience. Usable returns were received from 1048. Additional data were secured from the records of the several colleges.

25. KELLY, E. L., and WHITNEY, F. L. "Educational Magazines Read by Five Hundred Elementary School Principals and Classroom Teachers," *Elementary School Journal*, 29: 176-80, November, 1928.

Check lists of educational magazines were sent to elementary principals and teachers, members of the Department of Elementary School Principals. The respondents were located in 156 cities of all sizes in forty states and the District of Columbia.

26. KELLY, F. J. *The American Arts College, A Limited Survey*. New York: The Macmillan Company, 1925. 198 pp.

Data were collected by visits to four state universities, three endowed universities, five endowed colleges, and one city university. Brief visits were made to four other institutions. On these visits conferences were held with presidents, deans of the arts colleges, representative members of the faculties, and with student leaders. Additional data were secured from alumni by means of a questionnaire.

27. KLEIN, A. J., et al. "Survey of Land-Grant Colleges and Universities," *United States Office of Education Bulletin*, 1930, No. 9. Washington: Government Printing Office, 1930. Vol. I, 998 pp. Vol. II, 921 pp.

Questionnaires prepared by specialists with the aid of advisory com-

mittees were mailed to the faculties and former students of land-grant institutions. The magnitude of this survey investigation is indicated by the fact that 12,032 staff members and 37,342 former students filled out questionnaires. The care used in tabulating and organizing the data is indicated by the fact that "When errors, discrepancies, and omissions were discovered by the specialists . . . the pages in question were returned to the institution with a request for correction or explanation."

28. LONGSHORE, W. T., et al. "The Elementary School Principalship," *The Seventh Yearbook of the Department of Elementary School Principals*. Washington: National Education Association, 1928, pp. 132-638.

Data used in this study were secured from previous research and from the administration of several questionnaires. Case studies were made of a number of outstanding principals.

29. LUNDEEN, G. E., and CALDWELL, O. W. "A Study of Unfounded Beliefs among High School Seniors," *Journal of Educational Research*, 22: 257-73, November, 1930.

A list of 200 unfounded beliefs of wide geographical distribution was compiled in the form of a questionnaire to which students could respond with information respecting whether or not they had heard, believed, or were influenced by the beliefs. One thousand thirty questionnaires were filled out by high school seniors in ten high schools. Similar data were secured for purposes of comparison from 294 college students in three colleges.

30. MCGAUGHY, J. R. "The Fiscal Administration of City School Systems," *The Educational Finance Inquiry*, Vol. 5. New York: The Macmillan Company, 1924. 95 pp.

Data were collected with respect to 377 widely distributed cities from documents of the United States Office of Education, of the National Committee for Chamber of Commerce Coöperation with the Public Schools, and from the *Commercial and Financial Chronicle*.

31. MCINTOSH, H. W., and SCHRAMMEL, H. E. "A Comparison of the Achievement of Eighth-Grade Pupils in Rural Schools and in Graded Schools," *Elementary School Journal*, 31: 301-06, December, 1930.

Distributions of scores of 1921 pupils in graded schools and 1611 pupils in rural schools are presented in this study. Comparisons are made between the measures of central tendency and variability of these two groups of pupils.

32. MAXWELL, C. R., et al. "A Report on College Freshmen for the

First Semester of 1928-29," *North Central Association Quarterly*, 4: 484-600, March, 1930.

A questionnaire was sent to the secondary schools accredited by the association requesting the names of graduates and the colleges in which they enrolled. A questionnaire was then sent to the colleges with the names of reported entrants listed. This questionnaire requested information with respect to each entrant's success or failure in several college subjects.

33. MELCHIOR, W. T. "Insuring Public School Property," *Teachers College, Columbia University Contributions to Education*, No. 168. New York: Bureau of Publications, Columbia University, 1925. 187 pp.

A complicated and lengthy questionnaire was sent to every school district of New York State under the sponsorship of the State Department of Education. A 57 per cent response was secured and questionable data were verified by further inquiry.

34. MONROE, W. S. "A Survey of the Requirements for the Doctor of Philosophy in Education," *School and Society*, 31: 655-61, May 17, 1930.

Data were secured by questionnaire from each of the 29 institutions granting the doctor's degree in education more than once in the decade 1920-30. Findings were validated by having the respondents read critically a preliminary report of the study.

35. NELSON, M. G. "Subject Combinations in the Programs of Teachers in Small Secondary Schools in New York State," *School Review*, 37: 426-32, June, 1929.

Data were secured from teaching schedules obtained from 210 of the 350 small secondary schools to which requests were sent.

36. ODELL, C. W. "The Progress and Elimination of School Children in Illinois," *University of Illinois Bulletin*, Vol. 21, No. 38, *Bureau of Educational Research Bulletin*, No. 19. Urbana: University of Illinois, 1924. 76 pp.

The progress record blanks used in collecting data included such items as grade entered, grade at present, number of times failed of promotion, and number of times skipped. Usable records were secured with respect to 53,000 urban elementary pupils, 5500 rural elementary pupils, and 8500 high school pupils.

37. OJEMANN, R. H. "The Constant and Variable Occupations of the United States in 1920," *University of Illinois Bulletin*, Vol. 24, No. 39, *Bureau of Educational Research Bulletin*, No. 35. Urbana: University of Illinois, 1927. 47 pp.

Data were collected from United States Census Reports of 1920. The

findings were compared with those reported in 1914 by Ayres who used the 1900 Census Reports.

38. PRICE, R. R. "The Financial Support of State Universities," *Harvard Studies in Education*, Vol. 6. Cambridge: Harvard University Press, 1924. 205 pp.

Data were secured from reports and compilations of Federal and State legislation; bulletins of the U. S. Bureau of Education and other Federal departments; annual or biennial reports of state officers, such as auditors, tax reports, such as those of boards of regents, presidents, and finance offices, also annual catalogs; histories of the institutions; surveys of the institutions; general books of reference or textbooks containing compilations of pertinent data; published proceedings and transactions of associations of universities and colleges. The investigator presents a critical review of the sources used with particular reference to their reliability.

39. REMMERS, H. H., and GRANT, A. "The Vocabulary Load of Certain Secondary School Mathematics Textbooks," *Journal of Educational Research*, 18: 203-10, October, 1928.

Twelve mathematics texts were analyzed with respect to vocabulary by taking one line on each of a selected sample of pages until a total of 1000 running words was included. Thorndike's list of 10,000 most commonly used words was taken as a criterion for identifying the technical vocabulary. A second count on a similar basis showed differences that may not be considered statistically significant.

40. RUFI, JOHN. "The Small High School," *Teachers College, Columbia University Contributions to Education*, No. 236. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 145 pp.

Five small high schools were studied in detail. Records were examined, tests were administered, classroom teaching observed, teachers interviewed, and library and laboratory facilities inspected. In addition, the communities in which the schools were located were studied.

41. SAVAGE, H. J., et al. "American College Athletics," *The Carnegie Foundation for the Advancement of Teaching*, Bulletin No. 23. New York: The Carnegie Foundation for the Advancement of Teaching, 1929. 383 pp.

The data collected in this inquiry were obtained by means of interviews. Five members of the inquiry's staff visited 130 institutions and consulted hundreds of students, teachers, alumni, and other persons. In some cases two or three visits were paid to an institution.

42. SCHWEGLER, R. A., and WINN, EDITH. "A Comparative Study of the

Intelligence of White and Colored Children," *Journal of Educational Research*, 2: 838-48, December, 1920.

In this study the Stanford-Binet was administered to 34 colored girls and 24 colored boys and to equal numbers of white children drawn at random from the same school.

43. SELKE, ERICH. "A Comparative Study of the Vocabularies of Twelve Beginning Books in Reading," *Journal of Educational Research*, 22: 369-74, December, 1930.

Beginning books of 12 series of readers were subjected to vocabulary analysis in this study. "Each word in each book was listed and its frequency determined."

44. SMITH, H. L. "A Survey of a Public School System," *Teachers College, Columbia University Contributions to Education*, No. 82. New York: Bureau of Publications, Teachers College, Columbia University, 1917. 304 pp.

This study constitutes a survey of the public schools of Bloomington, Indiana, in which data were collected during the years 1912-1915. More attention is given to measurement of pupil achievement than is characteristic of most surveys of this period. Several standardized tests were administered.

45. STUART, HUGH. "The Training of Modern Foreign Language Teachers in the United States," *Teachers College, Columbia University Contributions to Education*, No. 256. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 111 pp.

A questionnaire was mailed to 776 universities, liberal arts colleges, and teachers colleges and replies were received from 412. A second questionnaire on observation and practice teaching was mailed to 405 institutions and replies were received from 228.

46. STURTEVANT, S. M., and STRANG, RUTH. "A Personnel Study of Deans of Girls in High Schools," *Teachers College, Columbia University Contributions to Education*, No. 393. New York: Bureau of Publications, Teachers College, Columbia University, 1929. 150 pp.

Questionnaire, observation, and interview were used in collecting the data of this study. "Visits were made to five schools, in which the dean permitted the observer to sit in her office for a day, observing and recording her activities, and interviewing her when she was not busy with other people."

47. Terman, L. M., et al. "Mental and Physical Traits of a Thousand Gifted Children," *Genetic Studies of Genius*, Vol. I. Stanford, California: Stanford University Press, 1925. 648 pp.

A group of gifted children were carefully selected on the basis of intelligence tests as the final criterion. This group was investigated with respect to many characteristics including racial and social origin; intellectually superior relatives; health and physical history; school progress and intellectual history; play interests; reading interests; intellectual, social, and activity interests; character and personality traits. Comparisons are made with a group of average children. The report includes a re-survey of the gifted children two years after the collection of the initial data. A more recent volume reports the status of the group of children several years later, see

TERMAN, L. M., et al. "The Promise of Youth. (Follow-up Studies of a Thousand Gifted Children.)" *Genetic Studies of Genius*, Vol. III. Stanford, California: Stanford University Press, 1930. 508 pp.

48. THORNDIKE, E. L., and BREGMAN, E. O. "On the Form of Distribution of Intellect in the Ninth Grade," *Journal of Educational Research*, 10: 271-78, November, 1924.

The distribution of the intelligence test scores of over 14,000 ninth-grade pupils is given and is shown to be approximately normal by means of Pearson's test for goodness of fit.

49. THORNDIKE, E. L., and ROBINSON, ELEANOR. "The Diversity of High School Students' Programs," *Teachers College Record*, 24: 111-21, March, 1923.

Tenth-grade pupils in ten school systems indicated in writing what subjects they were taking during the school year.

50. WEBB, P. E. "A Study of Geometric Abilities among Boys and Girls of Equal Mental Abilities," *Journal of Educational Research*, 15: 256-62, April, 1927.

Data were collected in this study by the administration under carefully controlled conditions of a standardized geometry test to 624 boys and 506 girls in four California high schools. The mental ages of these students as given by administration of the Terman Group Test of Mental Ability were secured from school records. All of the geometry test papers were objectively scored and were rechecked.

51. WHEELER, L. R. "A Comparative Study of the Physical Growth of Dull Children," *Journal of Educational Research*, 20: 273-82, November, 1929.
52. WHIPPLE, G. M. Sex Differences in Intelligence-Test Scores in the Elementary School," *Journal of Educational Research*, 15: 111-17, February, 1927.

Intelligence test scores from the administration of the National In-

telligence Tests, Scale A were secured from 2198 elementary school pupils. Similar data were secured by the administration of the Illinois General Intelligence Scale to 2501 pupils.

53. WHITNEY, F. L. "Teacher Demand and Supply in the Public Schools," *Colorado Teachers College Education Series*, No. 8. Greeley, Colorado: Colorado State Teachers College, 1930. 139 pp.

Data were collected by examination of records, reports, and by means of a questionnaire sent to superintendents.

54. WOODRING, M. N. "A Study of the Quality of English in Latin Translations," *Teachers College, Columbia University Contributions to Education*, No. 187. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 84 pp.

One hundred and fifty Latin examination books and the same number of English examination books from the College Entrance Examination Board were analyzed in this study. The Hudelson Scale was used in measuring the quality of the translations and the quality of composition in the English examination books.

55. WOODY, CLIFFORD. "The Arithmetical Backgrounds of Young Children," *Journal of Educational Psychology*, 24: 188-201, October, 1931.

An inventory test in arithmetic was administered to approximately 3000 kindergarten, first-grade, and second-grade pupils in 39 different school systems widely scattered throughout the United States.

CHAPTER IX

STUDYING THE EFFECT OF A SPECIFIED CHANGE IN A GIVEN CAUSE

General character of the problems. A typical problem is one in which a question is asked concerning the effect upon average pupil achievement of changing from one method of instruction to another. The specified change, however, may be in any factor that affects pupil achievement such as the length of class period, size of class, or textbook used. Other problems call for determining the effect of a specified change upon other traits or conditions. For example, an investigator may seek the effect of a curriculum change upon the proportion of children completing the twelfth grade. Sometimes the problem is stated in a form that does not make explicit the nature of the questions asked.

1. What is the effect of summer school attendance on scholastic achievement?
2. What is the value of moving pictures as visual aids in instruction?
3. What is the optimum size of class for the various school subjects?
4. What are the most effective grade placements of curriculum materials?
5. In what grade should the study of arithmetic begin?
6. To what extent does training in Latin transfer to other fields of study?

In the first of these statements the specified change in attendance is from no summer school attendance to attendance. In the second, "value" is to be interpreted as meaning effect, and the problem may be restated "What is the effect upon pupil achievement of introducing motion pictures as visual aids?" The third problem may be thought of as calling for the determination of the relative effects of various changes in size of

class and the identification of the size of class for which the achievement is a maximum. The interpretation of the fourth, fifth, and sixth problems is similar. In the seventh the specified change is from no training in Latin to training in the subject and the effect is achievement in some other field of study.

TABLE X. AVERAGE SPELLING ACHIEVEMENT AND MINUTES PER DAY DEVOTED TO SPELLING, SEVENTH GRADE
After Rice

AVERAGE SPELLING SCORE	MINUTES PER DAY DEVOTED TO SPELLING
86.5	60
80.0	40
84.0	30
78.0	30
77.2	30
76.0	30
76.7	25
84.9	20
84.0	20
80.6	20
79.5	20
72.8	20
81.1	15
90.0	12
75.3	10

Comparative survey method as a means of determining the effect of a specified change in a given cause. The comparative survey method of studying the effect of a specified change in a given cause may be illustrated by Rice's ¹ study in which he attempted to determine the effect upon spelling achievement of variations in the time allotment for teaching the subject. He surveyed the spelling achievement of the pupils in a number of cities by administering the same test to all groups. In the cities included in this survey the time allotment in the seventh grade varied from ten minutes per day to sixty minutes per day. The resulting average spelling scores are given in Table X. Although the lowest average spelling score in this table corresponds to

¹ Rice, J. M. *Scientific Management in Education*. New York: Hinds, Noble and Eldredge, 1912. Chapters V and VI. The report was first published in *The Forum*, April and June, 1897.

twenty minutes per day, the next lowest to ten minutes per day, and the next to the highest to sixty minutes per day, it is evident that the correspondence between "minutes per day" and "average score" is far from perfect. The average spelling scores for twenty minutes per day are on the whole slightly greater than those for thirty minutes per day. Hence, one might conclude that twenty minutes per day was the optimum length of period for spelling. Such a conclusion, however, would not have a very sound basis. An attempt to interpret the data of this table raises the question of the comparability of the several schools. Although information is not available, it is not unlikely that the schools differed in several respects which might affect the average spelling score of the pupils in the seventh grade. For example, it is likely that more incidental attention was given to spelling in those schools for which the time allotment was small than in those for which it was more generous. There may have been differences in the textbooks used. The methods of teaching may have varied. Possibly the quality of the pupil material may not have been the same in all cities. In view of the uncertainty in regard to the equivalence of such factors in the several schools, it is obvious that the findings cannot be considered dependable.

Rice's study illustrates the weakness of the comparative survey method of determining the effect of a specified change in a cause. In order to overcome the difficulties noted, it is necessary to select populations that are equivalent with respect to pertinent conditions and to plan the instruction that the pupils receive. When this is done, the research is called a *controlled experiment*.

An illustration of a controlled experiment.¹ In an attempt to ascertain the effect of systematic instruction in reading arithmetical problems upon problem solving ability, two groups of fifth-grade classes were selected in the public schools of Decatur,

¹ For a more extended account, see Monroe, W. S., and Engelhart, M. D. "The Effectiveness of Systematic Instruction in Reading Verbal Problems in Arithmetic," *The Elementary School Journal*, 33: 377-81, January, 1933.

Illinois. In making the selection, classes were chosen so that the two groups would be approximately equivalent with reference to both pupil material and teachers. More exact equivalence of pupil material was secured by administering an intelligence test and selecting pairs of pupils on the basis of IQ's. The equivalence of the groups thus formed was checked with reference to initial ability in both reading and arithmetic. As a means of checking upon the equivalence of the teachers of the two groups, each class was visited. No evidence was noted which indicated significant differences in the teaching ability of the two groups. It also appeared that the directions given to the teachers of the experimental classes were being followed and that the teachers of the other classes were continuing their usual plan of instruction which involved no systematic training in the reading of arithmetical problems.

In contrast with Rice's investigation, this study is characterized by an attempt to secure equivalence of pupil material, teachers, and other factors that might be expected to affect solving problem achievement. In other words, the investigation qualifies as a controlled experiment.

Definition of terms. Consideration of controlled experimentation will be facilitated by introducing certain terms. A *variable* designates a changing magnitude such as the total mileage reading of the speedometer on a motor car, a person's bank balance from day to day, his chronological age, the mean achievement of a class from month to month. The concept of a variable may be applied to a group of measures such as the scores of a class on a test, the salaries of a group of teachers, the size of classes within a school or group of schools, and the like. In such cases it is necessary to think of the measures as being arranged in a fixed order. Then the magnitude of the trait or characteristic will vary from measure to measure. A series of comparable textbooks or a series of comparable methods of teaching may also be thought of as a type of variable. In fact, any trait or characteristic in which a group exhibits individual differences may be thought of as a variable.

The concept of a variable implies continuity of change, which is illustrated by the total mileage readings of a speedometer. As the change is made from one reading to another, all intermediate values appear. A continuous variable may be measured to any degree of fineness that the instrument permits. This is true in the case of a person's height, or weight, or his chronological age. Size of a class must be measured in terms of integers. Hence, this variable is not continuous. But for many purposes it is not essential that a distinction be made between variables that are continuous and those that are discontinuous.

When the measures of a variable are in quantitative terms, their relative magnitude defines an order and the series is designated as an ordered one. Ratings in terms of verbal categories such as very poor, poor, average, superior, and excellent also form an ordered series. Some series, however, are unordered. A group of comparable textbooks illustrates an unordered series. Teaching methods, types of school organization, and occupations are other illustrations of an unordered series. The fact that a series is unordered, however, does not interfere with thinking of it as defining a variable.

A variable that is thought of as being contributed to by other variables operating as causes is designated as *dependent*. The causal variables are referred to as *independent variables* or as *factors* of the dependent variable. Hence, an experimental problem may be thought of as one of determining the effect upon the dependent variable of a specified change in a particular independent variable, commonly designated as the *experimental factor*.

A. PROCEDURE OF EXPERIMENTATION

The general plan of experimental research. The general plan of experimental research may be illustrated by considering a problem in which an investigator is seeking the effect upon pupil achievement of a specified change in some factor contributing to this dependent variable. When a group of pupils is subjected to instruction, their average achievement increases

with the passage of time. In order to secure a measure of the effect upon this growth in achievement of the specified change in an independent variable, it is necessary to determine the status that would have been attained if no change had been made. Hence, a controlled learning experiment requires the use of a minimum of two groups of pupils, one for each status of the experimental factor. Sometimes reference is made to single group experiments, but all that can be accomplished in such cases is the determination of the status of the dependent variable resulting from the functioning of the several independent variables. If a standardized test is used to measure the achievement specified as the dependent variable, the group to which the test was administered in the process of standardization may, under certain conditions, be utilized as the experimental group.¹

In a typical experiment two groups of pupils are selected so that they are equivalent with respect to the achievement designated as the dependent variable and with respect to all traits that may be expected to contribute to an increase in this achievement. Both groups are then subjected to the same instructional influences except that defined by the experimental factor. It is customary to apply the label *experimental group* to the one for which the status of the experimental factor is least typical. The other group is called the *control*. At the end of the experimental period, both groups are measured with respect to the dependent variable. The difference in gain or growth is the effect of the specified change in the experimental factor.

The dependent and independent variables of experimental problems in the field of education. The dependent variable is frequently the average achievement² of a group of pupils.

¹ For an illustration in which this was done, see Pratt, H. G., Dunlap, J. W., and Cureton, E. E. "The Subject-Matter Progress of Three Activity Schools in Hawaii, with a Note on Statistical Technique," *Journal of Educational Psychology*, 20: 494-99, October, 1929. The formula developed in this reference is incorrect. For the correct form, see Holzinger, K. J. "The Probable Error of a Difference Formula," *Journal of Educational Psychology*, 21: 63-64, January, 1930.

² Courtis has proposed that instead of taking pupil achievement as a depend-

This achievement may be limited to a narrow group of specific habits or it may include all outcomes of learning in a school subject or even in several subjects. Any effect, however, may be taken as the dependent variable. Variability of a group of pupils, average daily attendance, per cent of high school graduates entering college, efficiency of a school system, teachers' salaries, amount of reading done by pupils outside of school, and the like,¹ may be thought of as dependent variables.

Given a dependent variable, the independent variables are the several causes that contribute to this effect. We have only fragmentary information in regard to the identity of the causes contributing to the various effects in the field of education, but it appears that in some cases the number of independent variables may exceed twenty. When the effect is a segment of pupil achievement, the causes are the factors that affect learning. These include not only such variables as intelligence of pupils, length of class period, and other traits and characteristics that are measurable in quantitative terms, but also such factors as the textbook, method of teaching, and quality of supervision.

The experimental factor. Theoretically, any of the causes that contribute to the dependent variable may be studied experimentally to determine the effect of a specified change, but examination of a large number of representative experiments indicates that most experimental factors may be classified under a few heads. A large number of studies deal with a specified variation in the learning exercises the pupils are asked to respond to. A few typical variations are: (1) in teaching reading in the primary grades, phonics versus no phonics; (2) in a laboratory science, individual-laboratory exercises versus lecture-demonstration exercises; (3) in arithmetic, one type of

ent variable "change in the rate of growth" be used. This proposal is supported by theoretical arguments but certain difficulties are encountered in its practical application. Courtis, S. A. "The Measurement of the Effect of Teaching," *School and Society*, 28: 52-56, 84-88, July 14, July 21, 1928.

¹ See bibliography at end of chapter for illustrations of a variety of dependent variables.

practice materials versus another type of practice exercises; (4) intensive reading versus extensive reading; (5) in spelling, requests to learn and apply rules versus requests to repeat the spelling of words until memorization has been attained. Many learning exercises relate to certain materials of instruction. Hence, the nature of the request is changed when the materials of instruction are varied. In the case of drill, the number and distribution of the exercises may be varied.

Motivation procedures and techniques designate another experimental factor that has been studied by a number of research workers. This factor may be thought of as consisting of an unordered series of procedures and techniques designed to secure more intensive learning activity than would be stimulated by merely assigning learning exercises. Typical variations are: (1) use of interesting reading materials versus the use of materials judged to be less interesting, (2) short daily tests versus no such tests, (3) definite specific objectives versus general objectives, (4) information in regard to individual progress versus no such information, (5) group competition versus individual competition, (6) reproof versus commendation.

Other factors that have been studied experimentally are procedures for directing learning, diagnosis and remedial instruction, class size, length of class period, and classification of pupils.

Requirements for successful experimentation. As implied in the general description of experimental procedure on pages 274-75, the basic requirements for dependable findings are (1) selection of two or more equivalent groups of subjects, (2) maintenance of the specified status of the experimental factor in the experimental group and in the control group throughout the duration of the experiment, (3) control of the various non-experimental factors, (4) dependable measures of the dependent variable. Experimental studies will be successful to the extent that these requirements are met. A prerequisite for securing equivalent groups and for controlling the other non-experimental factors is identification of the pupil traits and other factors that contribute

to the dependent variable as defined by the problem. These traits and factors will vary with the nature of this variable, but a general consideration of the factors that contribute to pupil achievement will be helpful.

The factors contributing to pupil achievement. The research relative to the identification and potency of the factors that contribute to achievement is fragmentary and some of the findings do not appear to be highly dependable. However, the evidence relative to a number of factors appears convincing. As a means of convenience, the factors noted in the following pages are grouped under four heads.

- I. Pupil traits.
- II. Teacher factors.
- III. General school factors.
- IV. Extra-school factors.

I. *The significant pupil traits.* In a particular case we are concerned only with those pupil traits that affect the achievement specified by the problem. Obviously, such characteristics as color of hair, degree of beauty, and height would not be included in any case. On the other hand, general intelligence as measured in terms of a point score or of mental age, chronological age, and previous achievement in the field of experimentation would appear in the list. In some cases, study habits, attitudes, and interests and possibly other pupil traits should be included, but the evidence in regard to the contributions of such factors is very fragmentary and not entirely consistent.¹ Physical condition, except actual illness, seriously

¹For example, Symonds in concluding "a review of research on how to study" states that "the commonly accepted rules of study are often non-consequential." Symonds, P. M. "Methods of Investigation of Study Habits," *School and Society*, 24: 151, July 31, 1926.

For information relative to pupil traits see

Herriott, M. E. "Attitudes as Factors of Scholastic Success," *University of Illinois Bulletin*, Vol. 27, No. 2, *Bureau of Educational Research Bulletin*, No. 47. Urbana: University of Illinois, 1929, p. 31.

Gates, A. I. "A Study of Reading and Spelling with Special Reference to Disability," *Journal of Educational Research*, 6: 12-24, June, 1922.

Chambers, O. R. "Measurement of Personality Traits," *Research Adventures*

defective vision, or similar defects, and sex appear to be minor pupil factors.¹

1. There is abundant evidence that *general intelligence*, as measured by typical intelligence tests, influences the achievement of children. Many investigators have concluded that it is the most important factor.² Hence, general intelligence (mental age, or test score³) may be placed at the head of the list of significant characteristics of pupil material.

in University Teaching. Bloomington, Illinois: Public School Publishing Company, 1927, pp. 71-80.

Fleming, C. W. "A Detailed Analysis of Achievement in the High School," *Teachers College, Columbia University Contributions to Education*, No. 196. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 209 pp.

Fryer, Douglas. "Interest and Ability in Educational Guidance," *Journal of Educational Research*, 16: 27-39, June, 1927.

Ohmann, O. A. "A Study of the Causes of Scholastic Deficiencies in Engineering by the Individual Case Method," *University of Iowa Studies in Education*, Vol. 3, No. 7. Iowa City: University of Iowa, 1927. 58 pp.

Pressey, S. L. "An Attempt to Measure the Comparative Importance of General Intelligence and Certain Character Traits in Contributing to Success in School," *Elementary School Journal*, 21: 220-29, November, 1920.

¹ For example see

Hoefer, Carolyn, and Hardy, M. C. "The Influence of Improvement in Physical Condition on Intelligence and Educational Achievement," *Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1928, pp. 371-87.

Hall, Irene, and Crosby, Amy. "A Study of the Causes of Inferior Scholarship of Pupils in Low First Grade," *Journal of Educational Research*, 14: 375-83, December, 1926.

Mallory, J. N. "A Study of the Relation of Some Physical Defects to Achievement in the Elementary School," *George Peabody College for Teachers*, Contributions to Education, No. 9. Nashville: George Peabody College for Teachers, 1922. 78 pp.

Stalnaker, E. M., and Roller, R. D., Jr. "A Study of One Hundred Non-Promoted Children," *Journal of Educational Research*, 16: 265-70, November, 1927.

Westenberger, E. J. "A Study of the Influence of Physical Defects upon Intelligence and Achievement," *The Catholic University of America, Educational Research Bulletin*, Vol. 2, No. 9. Washington: The Catholic Education Press, 1927. 53 pp.

² For example, see Heilman, J. D. "Factors Determining Achievement and Grade Location," *The Pedagogical Seminary and Journal of Genetic Psychology*, 36: 435-57, September, 1929.

For a comprehensive account of the influence of general intelligence upon school achievement, see Terman, L. M., et al. "Nature and Nurture, Their Influence upon Achievement," *Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1928. 397 pp.

³ The IQ might have been listed as a pupil characteristic instead of mental

2. The significance of *chronological age* becomes apparent when a child having a mental age of twelve years and a chronological age of ten years is compared with one whose corresponding ages are twelve and fifteen. The first child has an IQ of 120 and the second one, an IQ of 80. Although the two children have equivalent mental ages, the first one is "bright" and the second is "dull." The significance of chronological age is further shown by a comparison of two children of the same IQ but of different chronological ages. Although the children are equally "bright," the difference in mental ages as well as the differences in physiological and social maturity emphasize the importance of chronological age as a factor contributing to school achievement. An excellent discussion of the influence of chronological age, or the maturity of which it is an index, has been provided by Commins.¹

3. *Previous achievement*² is a significant characteristic of the pupil material when it functions as a prerequisite for the learning involved in the experiment. For example, ability to read functions as a tool in learning arithmetic, geography, history, literature, and the like. Certain abilities in arithmetic and algebra function as tools in the study of chemistry, and achievement in chemistry may contribute to achievement in physics. Achievement in the first year of a foreign language functions as a tool in the more advanced study of that language. It would be easy to enumerate a large number of cases in which abilities engendered in a school subject function later in the learning of that subject or related subjects.

Abilities that function as a prerequisite for learning in one school subject may, or may not, be significant for learning in another school subject. For example, achievement in the first age, or intelligence test score. When both mental age, or intelligence test score, and chronological age are separately considered, the pupil is more adequately characterized and the IQ is superfluous.

¹ Commins, W. D. "Maturity and Education," *Educational Research Bulletin*, Vol. 3, No. 7, Catholic University of America. Washington: Catholic University Press, 1928, p. 36.

² The total outcome of learning includes general patterns of conduct as well as specific habits and knowledge. Among the possible outcomes are study habits which are not included here under the head of "previous achievement."

year of a given foreign language would be of more significance in an experiment in the second year of that language than it would be in an experiment in a different language. Achievement in the first year of a foreign language would probably be of negligible significance in an experiment that involved learning typewriting as the dependent variable. The previous achievement of children becomes of increasing importance as a factor in the achievement of the experiment in proportion to the extent to which the children have experienced subject-matter similar in content to that of the experiment.

II. *Teacher factors that affect pupil achievement.* Research has yielded little dependable information in regard to the teacher factors that contribute to pupil achievement.¹ Hence, any list of teacher factors must be recognized as a hypothesis. Amount of training, teaching experience, intellectual status, and personality are usually listed as important teacher factors, but they influence pupil achievement for the most part indirectly through their contributions to more immediate factors. Hence, the following list appears more useful.²

1. Instructional techniques:
 - a. Devising and assigning learning exercises,
 - b. Motivation procedures,
 - c. Directive procedures,
 - d. Diagnostic and remedial procedures.
2. Classroom-management procedures.
3. Skill in carrying out instructional techniques and classroom-management procedures.
4. Zeal of the teacher with reference to experimental factor.

¹Corey, S. M. "The Present State of Ignorance about Factors Effecting Teacher Success," *Educational Administration and Supervision*, 18: 481-90, October, 1932.

²The authors are aware of the widespread conviction that the personality of the teacher is an important educative factor, but such traits as breadth of interest, self-control, good judgment, leadership, forcefulness, honesty, adaptability, enthusiasm, and open-mindedness contribute to the teacher's instructional techniques and to his skill in the use of them and, hence, influence pupil achievement indirectly. The teacher's personality may make a direct contribution, but since experimental evidence is lacking and we do not have satisfactory means for measuring personality traits, it does not appear wise to include them in a list of teacher factors to be considered in experimental investigations. For a study

1. The attention given to methods courses in the professional training of teachers is evidence of a conviction that the *instructional techniques* employed by a teacher affect the achievement of pupils. This conviction is supported by some indirect evidence from investigations of the relation between the marks received by teachers in courses on methods of teaching and teaching success.¹ Since "instructional techniques" is a general designation, a classification under four captions is suggested: (a) devising and assigning learning exercises, (b) motivation procedures, (c) directive procedures, (d) diagnostic and remedial procedures. Recognition of these rubrics will enable an experimenter to be more certain in regard to the control of non-experimental factors under the general head of instructional techniques.

2. *Classroom-management procedures* include such items as taking the roll, distributing and collecting materials, starting the work of the period, dismissing the class in case the pupils go to another room at the end of the period, and dealing with disciplinary cases. The importance of these procedures is generally recognized. In fact, until recent years the teacher's ability as a disciplinarian was considered to be the most important of his qualifications. Although other aspects of teaching are now considered of more importance than the mere maintenance of order, adequate attention to routine matters of classroom management, inclusive of discipline, is regarded as essential for securing an environment that will facilitate learning. If, however, distinctly undesirable practices are avoided, it appears likely that variations in classroom-management procedures will not materially affect pupil achievement.

3. The effectiveness of an instructional technique or a classroom-management procedure depends upon the *skill* with which

of the relationship of certain traits to teaching success, see Morris, E. H. "Personal Traits and Success in Teaching," *Teachers College, Columbia University Contributions to Education*, No. 342. New York: Bureau of Publications, Teachers College, Columbia University, 1929. 75 pp.

¹ Knight, F. B. "Qualities Related to Success in Teaching," *Teachers College, Columbia University Contributions to Education*, No. 120. New York: Bureau of Publications, Teachers College, Columbia University, 1922, p. 42.

it is carried out. Although we have no means of obtaining precise measures of teaching skill, it is obvious that some teachers are more skillful in carrying out certain instructional techniques than are other teachers. When a new technique is being compared with a familiar one, it is likely that the new one will be applied less skillfully. For example, suppose an experiment is devised to determine the effect of supervised study in comparison with study without supervision. Suppose further that the plan of supervising study has been formulated in detail. If a teacher, who has become a skillful instructor under a plan that does not involve supervised study and who has not had experience in supervising study, attempts to teach one class employing supervised study and another without supervised study, it is reasonable to expect that he will be considerably more skillful in teaching the second class. If this is the case, the experiment would furnish a comparison between skillful teaching without supervised study and teaching with supervised study somewhat crudely carried out. Hence, the experiment would not yield satisfactory evidence of the relative merits of skillful teaching with supervised study and skillful teaching without supervised study.

It is difficult to demonstrate the influence of teaching skill because we have no satisfactory means for measuring this teacher factor. However, it not infrequently happens that variation in teaching skill appears the most probable explanation of relatively large gains in achievement. In a recent study of the effect of increasing the number of weekly tests in teaching spelling from two to three,¹ the findings for one grade group showed an extraordinary superiority of the two-test plan. In attempting an explanation of this result, the authors suggest teaching skill as an influential factor.

4. The *zeal* that a teacher exhibits in carrying out the instructional techniques he is employing is a subtle factor. It is related to the factor of skill, and perhaps the two overlap to

¹ Gates, A. I., and Bennett, C. C. "Two Tests versus Three Tests Weekly in Teaching Spelling," *Elementary School Journal*, 34: 44-49, September, 1933.

some extent, but there is evidence that indicates the presence of an important educative factor that differs in some respects from skill. The influence upon pupil achievement of the teacher's preference ¹ in regard to methods is indicated in the report ² of an experiment to determine the relative merits of instructional procedures that may be designated as Method A and Method B. Several teachers coöperated in the experiment, each teaching a class according to Method A and another class according to Method B. The following results were secured:

	NUMBER	MEAN SCORE	MEAN SCHOLASTIC GRADE
Pupils taught by Method A	417	71.5	83.9
Pupils taught by Method B	440	69.5	83.8
Gain in favor of Method A		2.0	

The teachers were asked to indicate which method they preferred. The following results were obtained when the data were tabulated according to the preference of the teachers:

Teachers Preferring
Method A

	NUMBER	MEAN SCORE	MEAN SCHOLASTIC GRADE
Pupils taught by Method A	131	75.0	84.8
Pupils taught by Method B	140	59.3	82.4
Gain in favor of Method A		15.7	

Teachers Preferring
Method B

	NUMBER	MEAN SCORE	MEAN SCHOLASTIC GRADE
Pupils taught by Method A	180	68.2	85.4
Pupils taught by Method B	178	72.2	85.2
Gain in favor of Method B		5.0	

¹ It is reasonable to expect that a teacher will exhibit greater zeal when employing a method that he believes in than when employing one that he does not like.

² Parr, R. M., and Spencer, M. A. "Should Laboratory or Recitation Have Precedence in the Teaching of High-School Chemistry," *Journal of Chemical Education*, 7: 571-86, March, 1930.

Teachers Having No
Preference

	NUMBER	MEAN SCORE	MEAN SCHOLASTIC GRADE
Pupils taught by Method A.....	80	67.0	82.7
Pupils taught by Method B.....	89	67.2	83.0
Gain in favor of Method B.....		0.2	

The differences between the mean scores of the several pairs of groups strongly suggest that the preference of the teachers in regard to the method of teaching affected the achievements of the pupils. If it is assumed that the preference in regard to methods affected the zeal of the teachers, it follows that this characteristic of teaching was an important educative factor. Several investigations ¹ contribute evidence in support of this conclusion.

III. *General school factors that affect pupil achievement.* Pupil achievement is affected directly or indirectly by several general school factors. For example, it is generally assumed that the textbook used in a course influences the achievement of the pupils. Much of this influence is indirect. For example, the character of the text influences the learning exercises assigned which in turn influence achievement. In the following list of general school factors no attempt is made to indicate whether a factor functions directly or indirectly.

1. Instructional materials (textbooks, library, maps, laboratory apparatus, etc.).
2. Time devoted to learning activity.
3. Concomitant training.
4. Size of class.
5. Administration and supervision.

¹Sexton, E. K., and Herron, J. S. "The Newark Phonics Experiment," *Elementary School Journal*, 28: 690-701, May, 1928.

Collings, Ellsworth. *An Experiment with a Project Curriculum*. New York: The Macmillan Company, 1923. 346 pp.

Pittman, M. S. *The Value of School Supervision*. Baltimore: Warwick and York, Inc., 1921. 129 pp.

Knight, F. B. "Qualities Related to Success in Teaching," *Teachers College, Columbia University Contributions to Education*, No. 120. New York: Bureau of Publications, Teachers College, Columbia University, 1922, p. 9.

1. *Instructional materials*, such as textbooks, library, and other school equipment, influence the learning activity of pupils through the learning exercises that they furnish or make possible. Texts in arithmetic, algebra, language, physics, and most of the other subjects furnish a number of learning exercises. Texts and other books make possible other learning exercises, such as requests to study certain pages or questions whose answers may be found by reading. In a similar way charts, maps, moving picture machines, laboratory apparatus, and the like, affect the number and type of learning exercises that may be assigned. Hence, the achievement of the pupils is likely to be affected by the instructional materials used with a class.

The intimate relation between instructional materials and learning exercises may make it impossible to have the former constant when the latter are greatly different. It should be noted, also, that certain types of learning exercises require certain instructional materials. Hence, if the purpose of an experiment is to compare two types of learning exercises, such as the demonstration lecture and individual laboratory work, the materials must differ. In such cases, the difference in instructional materials is essentially a phase of the experimental factor.

2. If the *time devoted to learning activity* is assumed to be an index of the amount of exercise of modifiable connections, it is apparently an important educative factor.¹ In group experiments the total time devoted to learning activity is affected by absences, but the length of the class period and the number of minutes per day devoted to study are more important unless there is a marked difference in the number of absences in the two groups. The time devoted to study should include that

¹ This statement appears to be defensible even though research has revealed a low correlation between school attendance and achievement.

Odell, C. W. "The Effect of Attendance upon School Achievement," *Journal of Educational Research*, 8: 422-32, December, 1923.

Denworth, K. M. "The Effect of Length of School Attendance upon Mental and Educational Ages," *Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1928, pp. 67-91.

denoted to thinking and talking about assignments as well as formal study, either at school or at home.

3. Pupil achievement in a given field may be affected by *concomitant training* in other fields. If two fields are closely related, what is learned in one may be an asset for learning in the other. Even when the fields are not closely related, there may be a transfer of study habits.¹

4. The *size of the class* disappears as an educative factor in an experiment where equivalent groups are secured by pairing, since this procedure secures classes of equal size. If the two groups are not equal in size, small differences do not appear to be significant because, within fairly wide limits, size of class does not appear to be an important educative factor.²

5. The *administration and supervision* of a school must be an important factor in learning activity if the attention given to these fields in teacher-training institutions is any criterion. However, it is difficult, if not impossible, to find any quantitative evidence in regard to the contribution of this factor to classroom learning. The reason for this seems to lie in the fact that any influence exerted by administration or supervision must be an indirect one operating through the teacher, the course of study, the organization of classes, the provision of school equipment, and the like.

IV. *Extra-school factors that affect pupil achievement.* Pupil achievement may be affected by several factors that have not been included in the preceding lists. The following appear to deserve consideration:

1. Participation in extra-class activities.
2. The pupil's home life.
3. Community interest in and attitude toward the school.

¹ For studies revealing transfer of this type, see

Gatchel, D. F. "Results of a How-to-Study Course Given in High School," *School Review*, 39: 123-29, February, 1931.

Hurd, A. W. *Problems of Science Teaching at the College Level*, Minneapolis: University of Minnesota Press, 1929. 195 pp.

Kornhauser, A. W. "Changes in the Information and Attitude of Students in an Economics Course," *Journal of Educational Research*, 22: 288-98, November, 1930.

² See reference on page 318 to the research by Hudelson on class size as an educative factor.

1. *Participation in extra-class activities* makes demands upon a pupil's time and when much time is devoted to such activities, the amount of learning outside of the class is likely to be affected. Under wise supervision, however, participation in extra-class activities may be beneficial to learning rather than detrimental.¹ Dramatic, scientific, technical, and debating clubs not only add interest to the school subjects to which they are related, but they may also contribute directly to achievement in certain fields.

2. The *child's home life* may influence his school achievement in many ways. Listening to conversation of parents and other members of the family, reading periodicals and books that the home affords, and traveling with members of the family are activities that may contribute to school achievement by providing a background of information for the learning that is to take place during the experiment. Topics in history, civics, biology, literature, and economics are more meaningful to the pupil who has had related experiences through travel. It is impossible to estimate the extent to which school achievement is influenced by these out-of-school experiences.

Parental supervision of home study probably affects school achievement,² but it is probable that the attitude of parents toward the school as an educative agency is a more potent influence than any supervision they may administer.

3. School achievement is influenced by *community interest in and attitude toward the school*. If the community is high in the socio-economic scale, the members of the community are likely to show much interest in school affairs and to coöperate with the principal and teachers in attaining the best conditions for school work. For example, the parents of such a community may coöperate with the school faculty in providing more adequate

¹ Monroe, W. S. "The Effect of Participation in Extra-Curricular Activities on Scholarship in the High School," *School Review*, 37: 747-52, December, 1929.

² Reavis, W. C. "Some Factors That Determine the Habits of Study of Grade Pupils," *Elementary School Teacher*, 12: 71-81, October, 1911.

Brooks, E. C. "The Value of Home Study under Parental Supervision," *Elementary School Journal*, 17: 184-94, November, 1916.

library facilities. In other cases, the community may be permeated with attitudes antagonistic toward the school and its administration. Such attitudes among parents tend to be acquired by pupils. Thus, community attitudes and interest may exert a subtle but possibly powerful influence on school learning.

Details of setting up and conducting controlled experiments.

Controlled experiments vary in details and hence what may be said relative to the technique to be followed in one experiment may not be wholly applicable to other experiments. The discussion in the following pages assumes an experimental group and a control group, each of which may involve two or more subgroups or classes. Except as explicitly indicated, the dependent variable is thought of as the average of a segment of pupil achievement and the experimental factor is one that affects learning. A person who attains an understanding of the procedures discussed should not encounter serious difficulty in making adaptations to problems that involve other dependent variables and other experimental factors. The discussion is organized under the following captions: (1) defining the problem, (2) securing a sample of school children representative of the population for which a conclusion is desired, (3) securing equivalent groups of pupils from this representative sample, (4) controlling the other non-experimental factors during the period of experimentation, (5) conducting the experiment, (6) measuring the dependent variable.

1. *Defining the problem.* In defining an experimental problem, attention should be given to five points: (1) nature and scope of the dependent variable, (2) nature and scope of experimental factor and the change to be made in it, (3) status of non-experimental factors, (4) duration of the application of the change in the experimental factor, (5) population for which a conclusion is desired.

The nature and scope of the dependent variable is frequently given only casual attention, but specification of these items is important. When the dependent variable is pupil achievement, it may be the sum of all types of outcomes of learning or it may

be limited only to certain ones.¹ For example, the outcomes of learning activity in the field of history include additions to the pupil's vocabulary, memorized facts, knowledge of events and their causes, points of view, attitudes, prejudices, interests, and study habits. In addition to the variations in definition suggested by this analysis, achievement in this field may differ with respect to topics studied (ground covered). The mere specification of "achievement in history" leaves the matter indefinite. If the change in the experimental factor is from heterogeneous to homogeneous grouping of pupils for instructional purposes, the dependent variable to be measured may be the amount of memorized information, total achievement, including attitudes and interests, average rate of progress (ground covered), or some other resultant of schooling. If the specified change is from the demonstration-lecture method to the individual-laboratory method, a number of dependent variables is possible, including the election of the science as a field of specialization. The choice of a test to measure the dependent variable implies specifications in regard to its nature and scope, but since indirect measurement is possible, the implications are uncertain. An experimental problem is not adequately defined until the nature and scope of the dependent variable are explicitly specified.

The experimental factor and the variation to be made in it must be defined with precision in order that the consequent change in the dependent variable may be ascribed to a definite cause. For example, a conclusion that Method A is superior to Method B is not very meaningful if the investigator defines these methods only by saying that they are the methods carried out in the experiment. In some cases, the nature of the experimental factor is such that it is relatively easy to give a precise definition of it. For example, Douglass compared the

¹ Although achievement is commonly thought of as outcomes of learning or the results of teaching, in most fields a large portion of what achievement tests measure is the same thing as is measured by general intelligence tests. Hence, strictly speaking, achievement tests do not yield pure measures of the results of teaching. This point is referred to again in considering the measurement of the dependent variable.

effectiveness of the recite-study sequence with the study-recite sequence where the divided period plan is used.¹ The reader of the report of the experiment has no difficulty in understanding the nature of the experimental factor and the change made in it.

When the experimental factor is complex, as is the case in many problems, it is difficult to plan the investigation so that the experimenter will be able to ascribe the change in the dependent variable to a specific cause. For example, if the assignment method is being compared with the project method, the change in the experimental factor involves so many phases that the interpretation of the findings cannot be made very specific. On the other hand, if the procedures to be followed in both the control and experimental groups are defined so that they differ in only one detail, abnormal or even unsound pedagogical conditions may be created and the conclusion will have only limited application. The dilemma suggested by these statements creates a serious difficulty in experimental research. Morrison has stated that it is "exceedingly difficult to raise an issue in the teaching process which is sufficiently definite."² When the experimental factor is complex, it may be possible to analyze the problem into a series of more restricted problems. The study of a single problem in this series will not have much practical significance and hence a person who plans a study of a complex problem should be prepared to deal with each of the restricted problems that the analysis may reveal.

The procedures defined by the experimental factor should have a common function, and this function should agree with the specifications of the dependent variable. Two methods of teaching may not have identical functions. For example, the lecture method may be, and probably usually is, directed toward engendering information and what is commonly designated as

¹ Douglass, H. R. "The Experimental Comparison of the Relative Effectiveness of Two Sequences in Supervised Study," *University of Oregon Publications*, Vol. 1, No. 4. Eugene: University of Oregon, 1927, pp. 173-218.

² Morrison, H. C. "The Major Lines of Experimentation in the Laboratory Schools," *Supplementary Educational Monographs*, No. 24. Chicago: University of Chicago Press, 1923, p. 5.

understanding. The class discussion method affords opportunity to engender ability to deal with thought questions, to develop ability in organizing arguments, and the like. Hence, these two methods may be conceived of in a form such that they have different functions. When this is the case, an experimental study of their relative effectiveness cannot be successful because the specification of a dependent variable compatible with one function will operate to the disadvantage of the other.

The point made in the preceding paragraph is important and recognition of the requirement of community of function of the procedures being compared would eliminate many ill-advised experiments. A critical examination of reports of experimental studies published since 1920 would reveal a surprising number in which the functions of the procedures compared are sufficiently dissimilar to make the inquiry not unlike a study of the relative effectiveness of two tools such as the hatchet and the saw.

An experimenter determines the effect of the specified change for a particular status of the several non-experimental factors. For example, a determination of the effect of a change in the number of minutes per day devoted to the teaching of spelling will be for pupils of a certain grade level and intellectual status, for a particular text, for a particular method of instruction, for a particular amount of incidental instruction, etc. Hence, in defining the problem there must be specifications relative to the status of the several non-experimental factors for which the determination is to be made.

The necessity of restricting the experimental factor indicates that a problem frequently analyzes into a series of related problems. The specification of the status of the non-experimental factors suggests further analysis. There is a problem for every variation in the status of each of the factors that functions as a cause of the dependent variable. Hence, the definition of a problem that appears to ask only a single question, may reveal an extended series of questions, each of which requires experimental study. The meaning of this statement will become more

apparent to the reader if he will attempt to define a few typical problems.

The effect sought is one resulting from the operation of the specified change in the experimental factor over a period of time. For example, in investigating the effect of changing the daily time allotment for the teaching of spelling from ten minutes to thirty minutes, we seek the effect of this change not for a single day or a single week, but for a semester or year or even a longer period. Hence, in defining the problem, the period during which the change is to be operative should be specified. Sometimes the effect of the change may be of such a nature that it is not observable until the change has been operative for a considerable period. For example, when homogeneous grouping versus heterogeneous grouping is made the basis of experimentation, it is conceivable that the effect upon pupil attitudes may not be apparent at the end of one semester, and the experimenter should be interested in the effect of this change in the grouping of pupils over a period of years. In many cases a change in method of teaching will not become fully effective until the pupils become acquainted with the new method and the teacher becomes skillful in its application.

The effect may be sought for a single pupil or a small group of pupils, but usually the researcher will be interested in the average effect for a large group or universe. Data are necessarily collected from a particular population, but we are interested in the determination primarily with reference to its application to other similar groups. In other words, we usually desire a generalized conclusion. Hence, the problem should be thought of as that of determining a generalized conclusion in regard to the average effect of a specified change in the experimental factor operating during a specified period and for a defined status of the other non-experimental factors.

2. *Securing a sample of school children representative of the population for which a conclusion is desired.* Application of the findings for a given group to another group requires that the investigator secure a group of school children that is sufficiently

representative to justify the generalization. One means of approximating representative sources of data in educational research is by the method of random sampling, but this technique is seldom feasible in experimental investigations, since the members of a random group would be scattered among the total population or universe and it would be difficult if not impossible to bring the several pupils together for the experimental instruction. Usually the investigator can only select a group that is judged to be representative.

It is contended by Lindquist ¹ that in many evaluations of instructional methods or materials some variation from representativeness of groups will not necessarily invalidate generalizations with respect to the effectiveness of the methods or materials compared. It is evident, however, that such variations cannot be permitted to be very great. Conclusions derived from groups of bright pupils may not safely be applied to average or dull pupils. The data obtained from dull pupils may not safely be used as a basis for generalizations which are to apply to average, or bright pupils. In the reports of the supervised study experiments of Breed ² and of Breslich ³ it is stated that the duller pupils profited most, while the bright pupils were not helped, and in some cases, were handicapped. If one of the experimenters had used only bright pupils and the other dull ones, the conclusions would be in opposition even though each may be a dependable basis for a generalization restricted to the population represented.

Large samples are likely to be more representative than small ones. Hence, an experimenter should endeavor to secure relatively large groups for his study. No minimum number of pupils can be specified. In an experiment to determine the effect of variations in size of class there should be several hundred pupils. In a study to determine the relative merits of two

¹ Lindquist, E. F. "The Standard Error of the Means of 'Matched' Samples," *Journal of Educational Psychology*, 22: 197-204, March, 1931.

² Breed, F. S. "Measured Results of Supervised Study," *School Review*, 27: 186-204, 262-84, March, April, 1919.

³ Breslich, E. R. "Teaching High School Pupils How to Study," *School Review*, 20: 505-15, October, 1912.

instructional procedures a smaller number may be satisfactory. In many cases two or more parallel experiments are more desirable than a single large group experiment. For example, if the resources of the experimenter permit including several hundred pupils in a study of the relative merits of two instructional procedures, a group of experiments should be planned rather than one large experiment.

3. *Securing equivalent groups of pupils.*¹ The control of pupil factors is accomplished by securing a control group that is equivalent to the experimental group. Complete equivalence is secured by matching the pupils in the two groups with reference to the traits that affect the dependent variable of the experiment. Precise matching with respect to more than one or two traits materially reduces the size of the groups because perfect mates cannot be found for many of the pupils. Hence, usually the matching is on the basis of one or two measures such as intelligence test score,² or initial achievement.³ Occasionally pupils are matched on the basis of a composite score.⁴ Olander⁵ paired pupils chiefly on the basis of their growth in arithmetical ability during a preliminary period in which all pupils were subjected to the same or similar instruction. Courtis⁶ has proposed a technique somewhat similar to that of Olander, but which is

¹ For a more comprehensive discussion of the techniques used in securing equivalent groups, see Engelhart, M. D. "Techniques Used in Securing Equivalent Groups," *Journal of Educational Research*, 22: 103-09, September, 1930.

² The following experiments are illustrations of this technique:

Anibel, F. G. "Comparative Effectiveness of the Lecture-Demonstration and the Individual-Laboratory Method," *Journal of Educational Research*, 13: 356, May, 1926.

Ullrich, O. A. "The Effect of Required Themes on Learning," *Journal of Educational Research*, 14: 296, November, 1926.

³ Burks, J. D., and Stone, C. R. "Relative Effectiveness of Two Different Plans of Training in Silent Reading," *Elementary School Journal*, 29: 433, February, 1929.

⁴ Douglass, H. R. "The Experimental Comparison of the Relative Effectiveness of Two Sequences in Supervised Study," *University of Oregon Publications*, Vol. 1, No. 4. Eugene: University of Oregon, 1927, pp. 173-218.

⁵ Olander, H. T. "Transfer of Learning in Simple Addition and Subtraction," *Elementary School Journal*, 31: 358-69, 427-37, January and February, 1931.

⁶ Courtis, S. A. "Criteria for Determining Equality of Groups," *School and Society*, 35: 874-78, June 25, 1932.

See also Courtis, S. A. "Maturation Units for the Measurement of Growth," *School and Society*, 30: 683-90, November 16, 1929.

more refined. If it is not feasible to classify the pupils into matched groups at the beginning of the experiment, matched groups may be selected when computing the mean gains. This procedure is likely to result in a material decrease in the number of cases but it has the advantage of being applicable to regular classes.

Several experimenters have considered the groups to be sufficiently equivalent when the means of the measures of the trait or traits were equal. Some have sought also equality of measures of variability.¹ Although the equivalence of the groups secured in these ways is not as precise as that obtained by means of matched pairs, it is likely that reasonably satisfactory control of pupil factors is attained, especially when the variabilities of the groups are considered. A practical advantage of the procedure is that the labor of handling the data is greatly reduced. When it is not feasible to secure equivalent groups, Melby and Lien ² have proposed the use of three or more regular classes without modification. Intelligence and achievement tests are administered to determine the initial status of these classes. The experimental procedure is then applied to the class whose initial status is superior to some of the classes, but is inferior to the others. If the final achievement of this class is superior to the final achievement of the initially superior class, or classes, then it may be argued with considerable justification that the method represented by the experimental procedure is relatively superior to that employed with the initially superior class. If, on the other hand, the final achievement of the class to which the experimental procedure was applied is below that of the initially inferior class, or classes, it may be argued that the method represented by this procedure is relatively inferior.

In a few specialized types of experiments the groups may be

¹ This technique is exemplified in Brooks, F. D. "The Transfer of Training in Relation to Intelligence," *Journal of Educational Psychology*, 15: 415, September, 1924.

² Melby, E. O., and Lien, Agnes. "A Practicable Technique for Determining the Relative Effectiveness of Different Methods of Teaching," *Journal of Educational Research*, 19: 255-59, April, 1929.

selected by a process of random sampling. In a study by the senior author¹ to ascertain how pupils solve arithmetical problems, four groups were secured by having four tests arranged in alternate order before they were distributed to pupils as they were seated in the classrooms. The four groups selected in this way were large and the assumption of equivalence seems to be justified.

Reeder² sought to secure equivalence by employing a rotation technique which involved interchanging the experimental and controlled groups at the middle of the experimental period. This procedure really divides the investigation into two sub-experiments. The data are organized so that the gains in achievement of both groups of pupils under the influence of the experimental procedure may be compared with the gains in achievement of both groups under the influence of control procedure. With reference to the total investigation this technique gives two groups which are equivalent in the sense that they are made up of the same pupils. However, the two groups are not necessarily equivalent when considered with reference to achievement in the field of experimentation, study habits, and possibly some other factors. If the problem is to determine the relative effects of procedures for directing study, a carry-over of study habits is to be expected in the case of the pupils who receive such instruction during the first half of the experimental period. A similar carry-over would, of course, be impossible in the case of the other group. Hence, this rotation technique would fail to secure groups equivalent with respect to study habits even though they consist of the same pupils.

The most precise equivalence will usually be attained when the groups consist of matched pairs, but when such a procedure is not feasible or is judged undesirable, one or the other tech-

¹ Monroe, W. S. "How Pupils Solve Problems in Arithmetic," *University of Illinois Bulletin*, Vol. 26, No. 23, *Bureau of Educational Research Bulletin*, No. 44. Urbana: University of Illinois, 1929. 31 pp.

² Reeder, E. H. "A Method of Directing Children's Study of Geography," *Teachers College, Columbia University Contributions to Education*, No. 193. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 98 pp.

niques may be employed. If the equivalence is to be judged only with reference to means and standard deviations, a procedure suggested by Rulon and Croon ¹ may be employed. When the groups of the experiment are not equivalent, it may be possible to calculate the effect of the lack of control of the pupil factors,² or, as suggested by Melty and Lien, dependable interpretation may be possible in spite of the lack of equivalence.

4. *Controlling the other non-experimental factors.* This phase of the experimental procedure introduces the assumptions that control of non-experimental factors is possible and that it can be accomplished without creating conditions which are significantly abnormal or pedagogically unsound. These assumptions appear reasonable in the case of a number of experimental problems, but in other cases their validity is not certain. When the experimental factor is related to certain non-experimental factors so that a change in the former requires changes in the latter, the assumptions can be satisfied only by analyzing the experimental factor and planning a series of experiments.

A non-experimental factor of considerable importance is that of the zeal or effort of the teacher.³ Preference for a method because of its novelty or because it is a current fad in education or because it is advocated by persons occupying positions of prominence is apt to stimulate a teacher to greater zeal in applying it than that with which the pupils of the other group are taught. Some degree of control of this subtle factor may be secured by carefully prepared instructions to the teachers, especially those having experimental groups, and by endeavoring to engender in the teachers a scientific attitude toward the investigation. A contribution to the control of the zeal of the

¹ Rulon, P. J., and Croon, C. W. "A Procedure for Balancing Parallel Groups," *Journal of Educational Psychology*, 24: 585-90, November, 1933.

² For illustrations see

Haefner, Ralph. "Casual Learning of Word Meanings," *Journal of Educational Research*, 25: 267-77, April-May, 1932.

Westfall, L. H. "A Study of Verbal Accompaniments to Educational Motion Pictures," *Teachers College, Columbia University Contributions to Education*, No. 617. New York: Bureau of Publications, Teachers College, Columbia University, 1934. 67 pp.

³ See pages 283-85.

teacher may be made by visiting the classroom, observing the teaching, and commenting to the teacher in a later conference on the extent to which her teaching appeared to be suitably zealous.

The skill of the teacher is another important non-experimental factor. If there is reason for believing that the teachers may not be equally skillful in instructing the experimental and control groups, a period of practice is desirable.

Frequently, control of non-experimental teacher factors is attempted by having the same teacher instruct both an experimental and a control group. The success of this technique is dependent upon the teacher being equally skillful and equally zealous in instructing the two groups. This condition may prevail but the teacher may prefer either the experimental procedure or the control procedure or carry them out with different degrees of skill. In any case a requisite for highly effective instruction is that the teacher believe that she is employing a good procedure. Hence, having an experimental group and a control group taught by the same teacher does not insure adequate control of teacher factors.

A method of rotation has also been employed as a means of neutralizing possible variations in teacher factors. According to this plan, at the middle of the experimental period a teacher who has been instructing an experimental group exchanges with one who has been teaching a control group.¹ Thus, it is argued that any difference in zeal or skill on the part of the two teachers will be neutralized by the fact that both the experimental group and the control group have received an equal amount of stimulation and direction from each teacher. This procedure will be successful in securing equivalence of these factors only when the two teachers are equally skillful and equally zealous in carrying out the procedures prescribed for the two groups. A teacher may be equally skillful and equally zealous in carrying out different procedures, but it appears likely that most teachers because of their lack of familiarity with or a dislike for one of the

¹ It should be noted that this exchange involves also a change of the teacher relative to the experimental factor.

procedures will teach with less skill and zeal in one of the groups than in the other. When this occurs the rotation procedure will not succeed in securing control of skill and zeal except by chance.

Another important non-experimental factor is the amount of time spent by the pupils in learning activity. The two groups of pupils should spend equal amounts of time in study and recitation. One of the best techniques that has been suggested for the control of learning time is that of providing teachers with detailed instructions in regard to assignments and the supervision of learning activity.¹ However, even wisely prepared instructions will not insure control of this factor, especially when the pupils are expected to participate in learning activity outside of the classroom. Furthermore, attempts to control learning time are likely to create artificial conditions that are in opposition to principles of good teaching.

When instructional techniques are being compared, it is important that the instructional materials be the same for both the experimental and control groups, and when instructional materials are being compared it is important that the instructional techniques be the same for both groups. Instructional techniques and materials of instruction, however, are closely related. It is impossible to have the latter constant when the former are greatly different. For example, certain types of learning exercises require certain types of materials of instruction. If the purpose of the experiment is to compare two types of learning exercises, such as demonstration-lectures and individual-laboratory work, the materials must differ. In such cases, the difference in materials of instruction is essentially a phase of the experimental factor.

5. *Conducting the experiment.* The preceding discussion of the control of non-experimental factors has included some reference to conducting the experiment. There are, however,

¹ For an excellent illustration of the use of such plans, see Coryell, N. G. "An Evaluation of Extensive and Intensive Teaching of Literature," *Teachers College, Columbia University Contributions to Education*, No. 275. New York: Bureau of Publications, Teachers College, Columbia University, 1927, p. 13.

certain points that should be emphasized. Wise planning is essential but not sufficient. The plans must be carried out. When several teachers are involved, the experimenter should prepare relatively detailed instructions for them to follow. If feasible, the teachers should be brought together for a conference at which the instructions are explained. Unless each teacher is instructing both an experimental group and a controlled group, separate conferences should be held with the two groups of teachers. In any case the instructions should be reduced to writing and a copy given to each teacher.

It is desirable that the investigator keep in close touch with the work throughout the duration of the experiment. Even when the instructions have been wisely formulated, conditions may arise such that strict conformity with them may not be compatible with good teaching. If possible, such cases should be reported to the investigator who should make the decision in regard to the procedure to be followed. As a means of obtaining a record of what occurs day by day, each teacher may be asked to keep a diary. If possible, the experimenter should visit each class a number of times. Information relative to details of instruction may be very useful in interpreting the experimental findings. For example, Brownell ¹ has reported an experiment in which the interpretation of certain findings was materially changed when certain instructional details were recalled.

6. *Measuring the dependent variable.* When an experimental problem is adequately defined, the nature and scope of the dependent variable will be specified. The research worker faces the task of selecting or devising an instrument to measure the specified variable, or more specifically the change in it. Usually this instrument will be an achievement test. In selecting or devising a test for use in an experimental study, testing time, convenience in scoring, and expense per pupil are minor considerations. Since the effect of variable errors upon a mean varies inversely as the square root of the number of cases, the

¹ Brownell, W. A. "An Evaluation of an Arithmetic 'Crutch,'" *Journal of Experimental Education*, 2: 5-34, September, 1933.

reliability of the instrument and its validity, as measured by the correlation of the test scores with a criterion, are of secondary importance. Systematic errors of measurement are controlled through the administration of the test and other aspects of testing conditions. Hence, the principal consideration in selecting or devising a measuring instrument for use in an experiment is that the systematic errors of validity be a minimum.

Current practices in experimental studies suggest that an objective test is essential. It is contended that an objective test measures the same thing as an essay examination within the same field ¹ and that the measures are more accurate measures because the scoring of the essay examination is subjective. Studies of the marking of examination papers exaggerated the unreliability of examination grades.² Furthermore, it has been shown that by exercising care in formulating the questions and by formulating rules for scoring, differences between the grades assigned by different teachers can be greatly reduced.³ Hence, the increase in reliability attained by using an objective test is much less than is commonly believed.

The evidence relative to the community of function of objective tests and essay examinations is not necessarily con-

¹ For example, see Wood, B. D. *Measurement in Higher Education*. Yonkers-on-Hudson: World Book Company, 1924.

Paterson, D. G. "Do New and Old Type Examinations Measure the Same Functions?" *School and Society*, 24: 246-48, August 21, 1926.

Corey, S. M. "The Correlation between New Type and Essay Examination Scores and the Relationship between Them and Intelligence as Measured by Army Alpha," *School and Society*, 32: 849-50, December, 1930.

Gilliland, A. R., and Misbach, L. E. "Relative Values of Objective and Essay Type Examinations in General Psychology," *Journal of Educational Psychology*, 24: 349-61, May, 1933.

² Monroe, Walter S., and Souders, Lloyd B. "The Present Status of Written Examinations and Suggestions for Their Improvement," *University of Illinois Bulletin*, Vol. 21, No. 13. *Bureau of Educational Research Bulletin*, No. 17. Urbana, University of Illinois, 1923. 77 pp.

³ Osburn, W. J. "Testing Thinking," *Journal of Educational Research*, 27: 401-11, February, 1934.

Peters, C. C., and Martz, H. B. "A Study of the Validity of Various Types of Examinations," *School and Society*, 33: 336-38, March 7, 1931.

Sims, V. M. "Improving the Measuring Qualities of an Essay Examination," *Journal of Educational Research*, 27: 20-31, September, 1933.

Stalnaker, J. M., and R. C. "Reliable Reading of Essay Tests," *School Review*, 42: 599-605, October, 1933.

vincing,¹ especially when the requirements of an experiment are considered. In such studies the purpose is to secure a measure of the change in achievement rather than of the status of achievement, and a test that is highly valid for measuring the latter may be a poor instrument for measuring the change in achievement.² Wiedemann and Neivens³ reported that a true-false test measures approximately 60 per cent of the same thing as a "compare-and-contrast" essay test.⁴ In view of the fact that in most fields achievement as measured includes what we call general intelligence as a large factor, it appears that the community of function of these two tests exclusive of this factor is materially less. Hence, when considered with reference to change in achievement, the two types of instrument should not be accepted as measuring the same thing. The change in achievement resulting from instruction will be in the factor that is not general intelligence. Hence the difference between mean score at the beginning of an experiment and that at the end is heavily weighted by a factor that is not influenced by instruction. At best the difference between the mean scores probably minimizes the growth due to instruction. The measurement by means of an objective test is likely to be indirect, and it is this condition that creates the possibility of errors of validity. Measurement by means of an essay examination can be and usually is much more direct. Hence, it seems justifiable to conclude that in many cases an essay

¹ For a strong argument on this point, see Cason, Hulsey. "The Essay Examination and the New Type Test," *School and Society*, 34: 413-18, September 26, 1931.

² For some evidence on this point, see Watson, Goodwin. "Note on Validity in the Measurement of Change," *Journal of Educational Research*, 27: 187-92, November, 1933.

³ Wiedemann, C. C., and Neivens, L. F. "Does the 'Compare-and-Contrast' Essay Test Measure the Same Mental Functions as the True False Test?" *Journal of General Psychology*, 9: 430-49, October, 1933.

⁴ A similar degree of community of function is reported for the "discuss" essay test and simple fact answer test and for the "explain" essay test and word-answer test.

Cochran, R. E., and Weidemann, C. C. "'Explain' Essay vs. Word-Answer Fact Test," *The Phi Delta Kappan*, 17: 59-61, December, 1934.

examination will be a superior instrument for measuring gain in achievement.

In devising a test for use in an experiment, an attempt should be made to formulate exercises that will call for the functioning of the abilities or traits specified as the dependent variable. It is relatively easy to approximate direct measurement of the more specific abilities such as motor skills and fixed associations. It is more difficult to secure satisfactory measures of knowledge achievement and generalized controls of conduct such as skill in reflective thinking, attitudes, ideals, and interests. The real test of a pupil's knowledge achievement is his ability to deal with difficulties and new situations. Hence, an instrument for the measurement of knowledge achievement should consist of thought questions. The test of general patterns of conduct is the consistency of conformity to the pattern. Hence, a single formal test cannot be expected to furnish adequate evidence of such achievement. The measurement of the acquisition of the study habits resulting from certain types of instruction has been attempted by measuring the knowledge achievement at the end of the experimental period. The real test of the acquisition of study habits is to be found in the conformity of the students to the procedures after the completion of the period of instruction.

Systematic errors of validity are due to fluctuations in the ratio of the mean of what is measured directly to the mean of the abilities or traits whose measurement is desired. Hence, when indirect measurement cannot be avoided, the experimenter should endeavor to select or devise an instrument such that this ratio will be the same for the control group as for the experimental group. This ratio will not be the same if the test favors either group. Usually it is not possible to secure objective evidence relative to this point and hence the experimenter must rely upon his judgment. A coefficient of validity cannot reveal the presence of a systematic error. Furthermore, it should be emphasized that a test or type of test that is shown to be satisfactory for one purpose is not necessarily equally

satisfactory in another situation. Usually the requirements of an experiment are highly specialized, and hence a test that is reported as satisfactory for a general survey of pupil achievement or for some other purpose may be a very poor instrument for measuring the dependent variable of an experiment.

Handling experimental data.¹ The usual method of handling the data obtained from a controlled experiment involving two groups involves no elaborate statistical procedures. The measurement of the dependent variable at the beginning and end of the experimental period yields two sets of data for each group of pupils. One procedure for handling these data is to calculate the following means:

M_{C1} = mean of first measures of control group

M_{C2} = mean of second measures of control group

M_{E1} = mean of first measures of experimental group

M_{E2} = mean of second measures of experimental group

The gain in the dependent variable is found by subtracting the mean of the first measure from that of the second.²

$G_C = M_{C2} - M_{C1}$ = mean gain of control group

$G_E = M_{E2} - M_{E1}$ = mean gain of experimental group³

The experimental difference, D , is obtained by subtracting G_C from G_E .

When the initial status of the dependent variable can be assumed to be zero, no test is administered at the beginning of the experimental period and

$$D = M_{E2} - M_{C2}$$

The experimental difference, D , if dependable, represents the effect of the specified change in the experimental factor.

¹ The plan described here is sometimes supplemented by the application of other techniques. Descriptions will be found in the illustrative references at the end of the chapter.

² It is necessary that the measures be comparable. The initial and final tests should represent equivalent forms. If they are not equivalent forms, but are equally valid with respect to the experimental achievement, conversion of the initial and final measures into standard scores, T-scores, age scores, or grade scores makes possible the calculation of mean gains. See pages 82 f.

³ These gains may also be obtained by calculating the gains of individual pupils and averaging them.

B. INTERPRETATION OF EXPERIMENTAL FINDINGS

Determining the dependability of an experimental difference by verification. An experimental difference is to be regarded as dependable when repetitions of the study yield similar differences. This statement suggests that an experimenter should endeavor to establish the dependability of his findings by repeating the study. Although this method may not often be feasible, it is to be recommended. Sometimes it is possible for the investigator to plan a group of similar experiments rather than a single large one. Barr ¹ has reported a suggestive illustration. A total of sixty-four subjects were involved in the study. Four similar experiments were conducted, each experimenter working with sixteen subjects. Comparison of the differences from the four experiments indicates their dependability. Although an experiment with as few as sixteen subjects is seldom to be commended, our confidence in the dependability of the reported differences is increased by the method employed.

Estimating the dependability of an experimental difference for the population of the experiment. When verification by repetition of the experiment is not feasible, a judgment in regard to the dependability of the obtained difference may be arrived at from a critical examination of the experimental procedure and the data collected. The reasoning involved is similar to that described for comparative surveys in the preceding chapter. In fact, an experiment is a comparative survey of selected populations which have been subjected to controlled educative influences.

The calculated difference represents the effect of the change made in the experimental factor plus the effects of the faults of the data and of failure to control completely the non-experimental factors. Hence, the problem is to determine whether the combined effect of imperfect control of the non-

¹ Barr, A. S. "A Study of the Amount of Agreement Found in the Results of Four Experimenters Employing the Same Experimental Technique in a Study of the Effects of Visual and Auditory Stimulation on Learning," *Journal of Educational Research*, 26: 35-45, September, 1932.

experimental factors and of data faults has been sufficient to give the obtained difference a sign opposite to that of the net difference. The obtained difference is regarded as dependable when it appears that the net difference would have the same sign. The reader should not confuse dependability with practical significance. An experimental difference is calculated from means and the limitations of averages are applicable. Furthermore, a very small net difference means that the change in the experimental factor has little effect.

The effect of failure to control non-experimental factors cannot be calculated, but an experienced investigator will usually be able to identify those non-experimental factors whose control is important in a particular experiment. The effect of inadequate control can only be estimated but usually it may be possible to make a dependable determination of its sign. General school factors and extra school factors should be critically examined for lack of control, but usually the most important variations in non-experimental factors are to be found in those that relate to the teacher. Teaching zeal and skill are especially difficult to control, and differences in them may make a significant contribution to the obtained difference.

Of the four types of error, variable errors of measurement and variable errors of validity are not likely to affect the difference very much, since the effect of these errors upon a mean is inversely proportional to the square root of the number of cases and a portion of this effect is likely to cancel out in the subtractions. When the test does not measure directly all phases of the specified pupil achievement, the systematic effect may be a matter of considerable importance. For example, experiments designed to determine the relative effectiveness of the individual-laboratory and lecture-demonstration methods of teaching a science have been criticized by pointing out that the test used did not measure directly some of the outcomes claimed for individual laboratory work, and hence favored the lecture-demonstration method.

If, as the result of the experimenters' consideration of the

control of non-experimental factors and data faults, it appears likely that the net difference has the same sign as that of the obtained difference and does not approximate zero, the calculated difference is designated as dependable. On the other hand, if it appears at all likely that the net difference has the opposite sign, the obtained difference must be labeled as lacking in dependability. Frequently it is not possible to make a dependable judgment with respect to the sign of the net difference. In such cases the dependability of the obtained difference should be regarded as uncertain.

Estimating the dependability of the obtained difference when considered with reference to a larger population or universe. The obtained difference is for the population included in the experiment. Usually we are interested in the net difference for a larger population or universe. If the experimental population is not representative, this condition may contribute to the obtained difference and hence affect its dependability when considered with reference to the larger population or universe. It is a common practice to calculate the probable error of the difference between the mean gains of two groups by means of the formula: ¹

$$PE_D = \sqrt{PE_{M_1}^2 + PE_{M_2}^2}$$

The result of the calculation is then compared with the obtained difference. McCall ² has proposed a plan of handling experimental data which culminates in the ratio $\frac{D}{2.78\sigma_D}$ which he calls the experimental coefficient (*EC*). The reader is then told that an experimental coefficient of 1.00 "means that we can be *practically certain* that the true difference is somewhere above zero." Although in another chapter McCall considers

¹ The corresponding formula for the standard error of the difference is

$$\sigma_D = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2}$$

Sometimes the long formula (see page 105) is used, but almost never with adequate recognition of the assumptions and implications involved in its use.

² McCall, W. A. *How to Experiment in Education*. New York: The Macmillan Company, 1923, pp. 140, 155.

the control of non-experimental factors, he gives the impression by this statement that when the experimental coefficient is found to be 1.00 or greater, the experimenter is justified in concluding that the *net* difference for the universe has been demonstrated to have the same sign as the obtained difference.¹

This procedure is subject to two criticisms. In the first place the use of the probable or standard error formulae implies that the experimental group and the control group are independent random samples of the universe. Random sampling is seldom feasible in selecting pupils for an experiment and if the two groups are chosen so they are equivalent, the samples cannot be independent. Lindquist and Wilks have proposed formulae for use when the groups have been matched.² The use of these formulae, however, is limited.

¹ The ratio $\frac{D}{PE_D}$, called the *critical ratio* (*CR*), is sometimes used. A value of 4.00 represents approximately the same situation as a value of 1.00 for the experimental coefficient.

² Wilks, S. S. "The Standard Error of the Means of Matched Samples," *Journal of Educational Psychology*, 22: 205-08, March, 1931.

Lindquist, E. F. "The Significance of a Difference between 'Matched' Groups," *Journal of Educational Psychology*, 22: 197-204, March, 1931.

See also

Ezekiel, Mordecai. "'Student's' Method for Measuring the Significance of a Difference between Matched Groups," *Journal of Educational Psychology*, 23: 446-50, September, 1932.

Lindquist, E. F. "A Further Note on the Significance of a Difference between the Means of Matched Groups," *Journal of Educational Psychology*, 24: 66-69, January, 1933.

Ezekiel, Mordecai. "Reply to Dr. Lindquist's 'Further Note' on Matched Groups," *Journal of Educational Psychology*, 24: 306-09, April, 1933.

Peters, C. C., and Van Voorhis, W. R. "A New Proof and Corrected Formulae for the Standard Error of a Mean and of a Standard Deviation," *Journal of Educational Psychology*, 24: 620-33, November, 1933. (In formulae (G) and (H) n should appear as \sqrt{n} .)

Monroe, W. S., and Engelhart, M. D. "A Critical Summary of Research Relating to the Teaching of Arithmetic," *University of Illinois Bulletin*, Vol. 29, No. 5, *Bureau of Educational Research Bulletin*, No. 58. Urbana: University of Illinois, 1931, pp. 100-07. Contains a description of the method proposed by Lindquist and Wilks. A standard error obtained by this method need not be regarded as a limit, as stated on page 105, if generalization is restricted to groups of the same distribution of initial measures.

Walker, H. M. "Concerning the Standard Error of a Difference," *Journal of Educational Psychology*, 20: 53-60, January, 1929. Use of the long formula for the difference between the mean gains of two equated groups is equivalent to use of "Student's" method. See both references to Ezekiel in this connection.

The second criticism is that the use of a probable error formula does not include any consideration of the effects of failure to control completely non-experimental factors or data faults except variable errors of measurement.¹ This is not a criticism of the procedure but rather of the practice of accepting proof of statistical significance as proof of the dependability of an experimental difference. The contribution to the obtained difference from failure to control completely non-experimental factors and from data faults, especially systematic errors of validity, is frequently relatively large and hence requires consideration. Since the assumption that the population of an experiment may be considered a random sample is probably seldom justifiable and since a demonstration of the statistical significance of the experimental difference is not proof of its dependability, experimenters should endeavor to establish the dependability of their findings as generalizations by other means. Proof of the statistical significance of an experimental difference is likely to be of minor importance.

The accomplishments of experimental research. Since controlled experimentation is generally recognized as a fruitful means of contributing to a science of education,² the accomplishments of this type of educational research will be commented on briefly. If experimental research in education is examined, a number of studies with dependable findings will be found. For example, the Judd-Buswell studies in the field of reading³ have added to our knowledge of the effect upon the mental processes of a reader when instructions or materials are varied. Studies relating to methods of teaching have added materially to our understanding of the processes of teaching and learning. The thesis that variations in teacher zeal and skill may be more

¹ It has been shown that the probable error formula for the effect of sampling includes also the effect of variable errors of measurement.

For proof see Huffaker, C. L., and Douglass, H. R. "On the Standard Errors of the Mean Due to Sampling and to Measurement," *Journal of Educational Psychology*, 19: 643-49, December, 1928.

² The progress toward a science of education is the topic of the concluding chapter of this volume.

³ One group of these studies was described briefly in Chapter I.

important than variations in instructional procedures may be cited as a major contribution. The total list of accomplishments of experimental studies, including the by-products, would doubtless be a long one and considering that we probably are only now emerging from the pioneer stages of this type of research we are justified in pointing to our accomplishments with considerable pride.

On the other hand, the relative number of dependable experimental studies is distressingly small. For example, in reporting a summary of the research relating to the methods of teaching mathematics at the secondary level, Douglass¹ states that the majority of more than two hundred studies examined are not worthy of mention. His summary includes only thirty-nine. Although it is likely that a number of the studies examined by Douglass were not controlled experiments, his statement is indicative of the quality of experimental research in this field. In the same number of the *Review of Educational Research*, Grinstead summarizes only eight experimental studies in the field of Latin and states that this list "includes all studies, known to be available, which make any significant contribution, however small, to Latin classroom method."² Summaries of experimental studies relating to a particular problem such as homogeneous grouping³ reveal inconsistencies in the several findings and usually the reviewer states that few, if any, conclusions may be regarded as definitely established.⁴

Sometimes an experimenter announces a conclusion so obvious

¹ Douglass, H. R. "Special Methods on the High School Level—Mathematics," *Review of Educational Research*, 2: 7, February, 1932.

² Grinstead, W. J. "Special Methods on High School Level—Latin," *Review of Educational Research*, 2: 56, February, 1932.

³ For example, see Rankin, P. T. "School Organization—Pupil Classification and Grouping," *Review of Educational Research*, 1: 215–29, June, 1931.

⁴ In a more recent review of the research relating to homogeneous grouping, Douglass states: "While homogeneous grouping has been frequently found more effective than heterogeneous grouping, the question has not been removed from that of unsolved problems, and the difficulties of adequate experimentation discourage hopes for any immediate definite answer."

Douglass, H. R. "Certain Aspects of the Problem of Where We Stand with Reference to the Practicability of Grouping," *Journal of Educational Research*, 26: 344–53, January, 1933.

that one wonders why the inquiry was ever attempted. In concluding, a summary of research on the effect of "special methods in techniques on comprehension," Gray ¹ states "the significant fact about all these studies is that comprehension usually increased when specific training to that end was provided." This conclusion appears to follow logically if we accept the thesis that children are teachable.

A detailed examination of the procedure of experimentation, such as that attempted in the preceding pages of this chapter, reveals a number of crucial difficulties and suggests that they cannot be overcome, at least in the case of many problems. In a recent article Brownell ² points out six faults that are frequently found in experimental studies. He also points out that experimentation in the field of education is not a simple undertaking and that many persons have attempted studies for which they were not properly equipped.

The total picture is thus one that should challenge research workers. Although the accomplishments are by no means negligible, they are certainly small in comparison with the total number of experimental studies. The time and money invested have yielded small returns. By way of explanation it may be pointed out that experimental inquiry in the field of education had scarcely begun by 1910 and that most of the work has been done since 1920. This explanation, however, is not adequate. When considered in the abstract, controlled experimentation seems to promise much. It is the procedure by means of which much has been accomplished in the field of the physical sciences. It is easily understood in its general outlines. However, when its application to the field of education is considered in detail, it becomes an extremely complex method of research. The cause of educational research, especially experimental research, has suffered from the enthusiasm of its friends. Shortly after 1920 there was a concerted effort among certain leaders to

¹ Gray, W. S. "Special Methods in the Elementary School—Reading," *Review of Educational Research*, 1: 253, October, 1931.

² Brownell, W. A. "Some Neglected Safeguards in Controlled Group Experimentation," *Journal of Educational Research*, 27: 98-107, October, 1933.

stimulate quantity production of educational research and teachers and other school people who had only very limited training were encouraged to undertake experimental studies. Under such conditions it was inevitable that there would be a large number of studies possessing little or no merit.

The outlook for experimental research. Now that attention is being directed to the techniques of experimentation and the crucial difficulties are being pointed out, it seems reasonable to expect a material improvement in the quality of this type of research. In the first place there should be fewer ill-advised undertakings. The nature and scope of the dependent variable should be specified in defining the problem. In the past this has often received little or no attention. The problem has been thought of as being to determine the effect of a specified change in the experimental factor without attempting to specify the nature of the effect to be measured. For example, in the study of the effect of changing the grouping of pupils for instructional purposes from a heterogeneous plan to a homogeneous plan, practically no attention has been given to the nature of the pupil achievement to be measured. Is the evaluation of the two procedures to be based on the ground covered, extent of the skills and memorized information acquired, ability to respond to thought questions, pupil attitudes and interests, or some combination of these outcomes of learning? Obviously the problem is not adequately defined until the nature of the dependent variable has been specified.

We may also expect more attention to be given to the similarity of the function of the two procedures being studied. Obviously, procedures that have dissimilar functions are not suitable for experimental study. It is easy to recognize this condition when the functions are grossly dissimilar as in the case of a method of instruction in arithmetic designed to engender calculation skills and a second method designed to engender problem solving ability. It is, however, not always easy to identify dissimilarities of function. For example, is the function of a small class the same as that of a large class? Not necessarily,

especially if the experimental factor is considered to include instructional procedures that are adapted to the conditions created by the number of pupils to be instructed. The primary function of a large class may be to contribute to the engendering of information while that of a small class may be to contribute to the engendering of the ability to discuss problems and issues and to respond to other types of thought questions.

Specification of the nature of the dependent variable to be measured and consideration of the similarity of the functions of the procedures being compared will greatly reduce the number of ill-advised experimental studies. Adequate definition of the problem should also reduce the number of attempts to prove the obvious. There is also opportunity to increase the quality of the research by improving the experimental procedure at certain points. In many cases the duration of the experiment should be extended. More attention should be given to the control of non-experimental factors. An effort should be made to secure more valid measures of the dependent variable. Finally, the quality of experimental research may be increased by more intelligent interpretation of the findings. Determinations of the statistical significance of the experimental difference should be replaced by more appropriate considerations of dependability. When at all feasible, the findings should be subjected to the test of experimental verification. The experimenter should seek the explanation of his findings. Frequently, an explanation of why the results were obtained is more significant than the actual findings.

ILLUSTRATIVE EXPERIMENTAL STUDIES

The following references are not given as models of experimental procedure but rather as additional illustrations of techniques that have been employed. A number of the studies are laboratory experiments. In several cases, the dependent variable is not pupil achievement. A few studies of transfer of training have been included.

BARR, A. S., and PARK, J. S. "An Experimental Study of Functional Learning," *Journal of Experimental Education*, 1: 9-17, September, 1932.

Two methods of studying were employed by groups of graduate students. In the first, or direct, method, the students were instructed to memorize the symbols of two artificial alphabets. In the second, or incidental method, the students were told to concentrate on the translation of meaningful material. A rotation technique was used in which efforts were made to measure and allow for practice effect. Objective tests, devised by the experimenters, were used to measure immediate and delayed retention. In addition to standard treatment of the data, learning curves were constructed. Additional data are reported with respect to the effects of attitude, effort, and fatigue. The experimenters are to be commended for the following statement: "The results are true only for the subjects, methods, materials, conditions, and the learning measured in this experiment." Generalization to ordinary school learning would, of course, be unwarranted.

BERGMAN, W. G., and VREELAND, WENDELL. "Comparative Achievement in Word Recognition under Two Methods of Teaching Beginning Reading," *Elementary School Journal*, 32: 605-16, April, 1932.

The "visual method" and the "picture-story method" of teaching beginning reading, compared in this experiment, are described in detail. Three pairs of schools were matched with respect to nationality and socioeconomic status of pupil populations and school organization. The teachers participating in the experiment were matched with respect to principals' ratings and "carefully devised regulations were put in force covering the length of class period, outside practice in reading by pupils, and the preparation of supplementary instructional materials." The pupils were tested four times during the experiment. Vocabulary analyses are presented of the contrasted reading materials and an analysis is given of the "relative fairness" of the final forms of the tests with respect to the compared reading methods.

BROWN, A. E. "The Effectiveness of Large Classes at the College Level: An Experimental Study Involving the Size Variable and the Size-Procedure Variable," *University of Iowa Studies in Education*, Vol. 7, No. 3. Iowa City, Iowa: University of Iowa, 1932. 66 pp.

This experimenter sought "to measure the value of a set of procedures believed to be suitable to a large group, by comparing the achievement of a large group using these procedures with that of a small class taught by the instructor's usual type of instructional procedure." The procedures are described in detail. Lindquist and Wilks' formula was used in calculating the standard errors reported in the study. Included in the report are "interpretations based on observational evidence" and student opinion on the new procedures and on large classes as obtained by means of a questionnaire.

BUSWELL, G. T., and JOHN, LENORE. "Diagnostic Studies in Arithmetic," *Supplementary Educational Monographs*, No. 30. Chicago: University of Chicago Press, 1926. 212 pp.

Two laboratory studies and two single group experiments under school conditions are reported in this monograph. In the first laboratory study, eye movements in column addition were studied. In the second, time analyses were made of the four fundamental operations.

CATTELL, PSYCHE. "Constant Changes in the Stanford-Binet IQ," *Journal of Educational Psychology*, 22: 544-60, October, 1931.

In experimentation of which this study is typical, the dependent variable is the IQ and the independent variable, or "experimental factor" is time.

CHARTERS, W. W. *Motion Pictures and Youth, A Summary*. New York: The Macmillan Company, 1933. 66 pp.

In this reference is summarized a series of coördinated studies which fall "into two groups: one, to measure the effect of motion pictures as such upon children and youth; the other, to study current motion-picture content and children's attendance at commercial movie-theaters to see what they come in contact with when they attend them." Studies of the first group classify as experiments in which the experimental factor is "current commercial motion pictures" and the dependent variables studied are "information, attitudes, emotions, health, and conduct." Some of the experimentation was of the laboratory type.

CLARK, MILDRED, and WORCESTER, D. A. "A Comparison of the Results Obtained from the Teaching of Shorthand by the Word Unit Method and the Sentence Unit Method," *Journal of Educational Psychology*, 23: 122-31, February, 1932.

One hundred and nine pupils in six high schools were taught by the sentence unit method and 83 pupils in five high schools were taught by the word unit method. Comparisons were made between the achievements of the two groups on several tests and also between the achievements of two groups of 44 pupils each paired from the entire group on the basis of intelligence test scores and chronological age.

COLLINGS, ELLSWORTH. *An Experiment with a Project Curriculum*. New York: The Macmillan Company, 1923. 346 pp.

The experimental factor in this investigation was essentially that of variation in the type of learning exercise. The experimental group of 41 rural pupils proposed the learning exercises themselves, while the control group of 60 rural pupils had the traditional type of learning exercises assigned to them. At the end of four years, achievement was measured by a number of standardized tests. Data are presented with respect to outcomes

other than those measured by the tests. The conclusions are favorable to the project type of learning exercise, but the experimenter may be criticized for failure to control important non-experimental factors including the zeal of the teachers.

DOUGLASS, H. R. "The Experimental Comparison of the Relative Effectiveness of Two Sequences in Supervised Study," *University of Oregon Publications*, Vol. 1, No. 4. Eugene: University of Oregon, 1927, pp. 173-218. For a briefer account see: Douglass, H. R. "An Experimental Investigation of the Relative Effectiveness of Two Plans of Supervised Study," *Journal of Educational Research*, 18: 239-45, October, 1928.

The problem was to determine the relative effectiveness of the study-recite sequence in supervised study as compared with the recite-study sequence. Ten pairs of groups averaging 14 pupils each, were selected and carefully equated on the basis of age and a composite of regressed, or estimated true, intelligence test scores and achievement test scores. To equate teacher factors and those of room environment, teachers and rooms were exchanged at the mid-point of the experiment. At the end of eleven weeks, the final achievement tests were administered. In interpreting the differences in mean gains, the long formula for the standard error of a difference was used. Douglass also presents information with respect to the teachers' attitudes toward the experimental factor and the relation of the experimental factor to variability in achievement.

DYNES, J. J. "Comparison of Two Methods of Studying History," *Journal of Experimental Education*, 1: 42-45, September, 1932.

In studying social science materials, one group read and reread the material while the pupils of the other group gave the material a rapid reading; reread the material, underlining essential parts, and taking notes; and recalled what was read. A rotation technique was employed with 144 pupils in two high schools. Each set of study material was used half of the time with one method and half of the time with the other. A test was given before and after study of the material, and retention was tested three weeks later, a very commendable procedure. The following statements reveal a somewhat unique method of handling data ". . . of the 144 pupils who participated in the final experiments, 67 learned more with Method X, 73 with Method Y, and 4 pupils made equal gains with both methods. For retention, Method X proved to be the better for 49 pupils and Method Y was better for 76, while 9 pupils retained equal amounts with either method." This method of handling data seems more meaningful than the "statistically significant" differences reported.

EATON, M. T. "The Effect of Praise, Reproof, and Exercise upon Muscular Steadiness," *Journal of Experimental Education*, 2: 44-59, September, 1933.

The apparatus used to measure the effect of praise, reproof, and exercise upon muscular steadiness was a plate-and-stylus tester of the Whipple type. The subject held the stylus horizontally at arm's length and placed it in the hole in the vertical contact plate. The operator "pressed the key that connected the stylus and contact plate with the electric counter and simultaneously started the stop watch . . . only the net number of contacts in the specified ten second testing period was recorded." The subjects were tested several times in different types of stimulus situations. The responses are analyzed in detail.

HORTON, R. E. "Measurable Outcomes of Individual Laboratory Work in High School Chemistry," *Teachers College, Columbia University Contributions to Education*, No. 303. New York: Bureau of Publications, Teachers College, Columbia University, 1928. 105 pp.

Several experimental investigations of the individual laboratory, lecture demonstration, and problem methods in high school chemistry are reported in this monograph. After preliminary experimentation, this experimenter set up nine groups, varying in size from 26 to 128 pupils, approximately equivalent with respect to means and standard deviations on the mid-term examination in chemistry. Horton's experimentation is commendable in several respects: comparatively large groups were used, experimental factors were precisely defined, the groups were probably of adequate equivalence, precautions were taken to secure control of important non-experimental factors, and efforts were made to measure a variety of outcomes.

HUDELSON, EARL. *Class Size at the College Level*. Minneapolis: University of Minnesota Press, 1928. 299 pp.

Data were collected with respect to the attitudes of students and teachers toward class size by means of questionnaires and information relative to the instructional techniques employed by teachers, reputed to be skillful with large classes, was collected by observation. Trends in class size were investigated and the costs of large and small classes compared. The effects of class size, as measured by term marks, was investigated. An extensive program of experimentation was carried on which involved 6059 students in 104 classes taught by 21 instructors.

HURLOCK, E. B. "An Evaluation of Certain Incentives Used in School Work," *Journal of Educational Psychology*, 16: 145-59, March, 1925.

Four groups of elementary school pupils were used. After equating with respect to several factors, the members of the first group were praised in

the presence of their classmates with respect to their achievement on the initial test, the members of the second group were reproved, the members of the third group heard this praise and reproof, and the fourth group was taught separately. This procedure was repeated. The following differences in achievement are reported: praised-control, reproved-control, ignored-control. These differences are accompanied by their probable errors. The data are analyzed to show the relation of the effects of the motivating factors to age, sex, initial ability, and accuracy.

JUDD, C. H., and BUSWELL, G. T. "Silent Reading: A Study of the Various Types," *Supplementary Educational Monographs*, No. 23. Chicago: University of Chicago Press, 1922. 160 pp.

In this laboratory study, the effects on eye movements of changes in content of reading were studied. These changes included changes in difficulty, changes in language, and changes in attitude or purpose.

KNOWLTON, D. C., and TILTON, J. W. *Motion Pictures in History Teaching*. New Haven: Yale University Press, 1929. 182 pp.

In this experiment, photoplays supplemented the instruction in American history of the experimental group. The pupils of the control group were given supplementary pages containing the information presented in the photoplays but not in the regular text. This experiment is commendable in that efforts were made to measure a variety of outcomes. This is indicated in the following section titles: "The Effect of Photoplays upon Retention" and "Comparison of Experimental and Control Groups as to Participation in Classroom Discussion, Expression of Interest, and Voluntary Reading."

LEONARD, J. P. "The Use of Practice Exercises in the Teaching of Capitalization and Punctuation," *Journal of Educational Research*, 21: 186-90, March, 1930. See also Leonard, J. P. "The Use of Practice Exercises in the Teaching of Capitalization and Punctuation," *Teachers College, Columbia University Contributions to Education*, No. 372. New York: Bureau of Publications, Teachers College, Columbia University, 1930. 78 pp.

Eighty-two eighth- and ninth-grade pupils were paired on the basis of composite scores derived from the scores on several tests. "In addition to the regular mimeographed lesson sheets, the experimental group received special practice exercises in proofreading, error correction, and dictation." Twenty-five minutes of each period were devoted to these. The control group followed the more conventional method "such as picking out punctuation marks from paragraphs and citing rules for their use, writing compositions and reviews for books and plays, and formulating sentences to illustrate certain rules." Several tests were used to measure achievement.

Three compositions and two letters written by the pupil were carefully scored for errors. In interpreting the experimental differences, both the long and short formulae were used, but no allowance was made for selection resulting from pairing. (See page 309.)

MALLER, J. B. "Cooperation and Competition, an Experimental Study in Motivation," *Teachers College, Columbia University Contributions to Education*, No. 384. New York: Bureau of Publications, Teachers College, Columbia University, 1929. 176 pp.

Maller used 814 experimental and 724 control pupils in this investigation of individual and group competition as motivating factors. The experimental pupils, alternately stimulated by individual recognition and reward and by group or class recognition and reward, solved addition examples. The investigator states in this connection: "The motives of self and class were alternated six times, respectively. The problem of practice effect was thus practically eliminated. All conditions of work aside from the motives were identical." The difference in favor of individual competition was almost thirteen times its probable error. There is little reason to doubt the "statistical" significance of this difference. The experimental conditions, however, may be characterized as abnormal. The effectiveness of competition would lessen with continued use.

NELSON, M. J. "The Differences in the Achievement of Elementary School Pupils before and after the Summer Vacation," *University of Wisconsin, Bureau of Educational Research Bulletin*, No. 10. Madison: University of Wisconsin, 1929. 48 pp.

The experimental factor and the dependent variable are evident in the title of this study. An additional factor which was studied in the case of arithmetic and spelling was the "time elapsed before pupils return to the Spring level of achievement."

OLANDER, H. T. "Transfer of Learning in Simple Addition and Subtraction," *Elementary School Journal*, 31: 358-69, 427-37, January, February, 1931.

This investigator used 300 pairs of second-grade pupils equivalent with respect to *growth* in arithmetic ability over a period of five weeks. For twelve weeks the pupils of the experimental group were given instructions in generalizing for three minutes of the daily twenty-minute period. Achievement was tested several times during the experiment. The criticism seems justifiable that the experimental factor, generalizing instruction was not applied for a sufficient time or intensively enough to add materially to the generalizing abilities acquired by the pupils on their own account. The experiment is unique in the technique used to secure equivalence of groups.

REEDER, E. H. "A Method of Directing Children's Study of Geography," *Teachers College, Columbia University Contributions to Education*, No. 193. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 98 pp.

In this rotation experiment in seventh-grade geography the study of the pupils during experimental learning differed from that of the pupils during control learning in that the former involved the use of mimeographed sheets of study questions with each assignment. While the rotation technique may be criticized with respect to carry-over of study habits from experimental learning to control learning, this limitation is one which operates to reduce differences in achievement. Reeder's findings are significantly favorable to the directive procedure which constituted the experimental factor in spite of this limitation.

RULON, P. J. *The Sound Motion Picture in Science Teaching*. Cambridge: Harvard University Press, 1933. 236 pp.

A study was made of the socio-economic levels of the communities in which the schools participating in this experiment were located. Other factors studied include the geographical distribution of the pupils, the organization of their schools, the types of enrollment in general science, the comparative teaching loads and proficiency of the teachers. The pupils were equated on the basis of intelligence test scores, achievement test scores, and chronological ages. Equivalence was checked with respect to geographical location, "occupational-status" scores, pupil-hour load of the teacher, size of class, and sex. The experimenter aided in the production of the sound films and text used in an effort to have them supplementary to each other and prepared the achievement tests used.

STARCH, DANIEL, and ELLIOTT, E. C. "Reliability of the Grading of High-School Work in English," *School Review*, 20: 442-57, September, 1912.

This well known pioneer study may be regarded as an experiment in which the "personal equation" is the experimental factor and the mark assigned to an examination paper the dependent variable. Facsimiles of two examination papers in English were rated independently by a number of teachers. In general, a change in the person rating the paper resulted in a change in the grade assigned.

WOOD, B. D., and FREEMAN, F. N. *An Experimental Study of the Educational Influences of the Typewriter in the Elementary School Classroom*. New York: The Macmillan Company, 1932. 214 pp.

Several thousand pupils participated in this experiment. The experimental and control teachers are shown to be approximately equivalent with respect to several traits or characteristics. The experimental pupils and the control pupils are shown to be equivalent with respect to CA, MA, IQ, and

initial achievement. A number of standardized tests were used in measuring achievement. The report contains interesting graphic presentations of data. The following titles indicate the scope of the research: "Comparative Gains in General Educational Achievement," "Comparative Gains in Individual Subject Matters," "Writing Done by Experimental and Control Children," "The Typewriter and the General Aims of Education," "The Typewriter and the Pupil's School Interests," and "The Typewriter and Reading."

CHAPTER X

STUDYING PROBLEMS OF PREDICTION

The problems of prediction.¹ Intelligence test scores and other prognostic measures are used as a basis for predicting school marks and future status in other lines of endeavor. Predictions from such variables may be made in several ways. A person may make estimates without conforming to any definite procedure. For example, in predicting success in college, a high school principal might study a student's record including intelligence test score, chronological age, occupation of father, participation in activities, and the like, giving to each item the weight he considers appropriate in the particular case. Predictions made in this way may be highly accurate, but the procedure is not systematic. Usually predictions are thought of as being made by means of some systematic procedure such as is represented by a formula.

The two general problems are those of seeking out the best prognostic measures for a given situation and deriving the most effective systematic plan or formula for making the desired predictions or estimates from them. In both cases, the purpose is to reduce the error of prediction or estimate to a minimum. Hence, the development of techniques for determining the magnitude of errors of estimate creates a subordinate problem of major importance.

A. METHODS OF MAKING PREDICTIONS

Types of prediction formulae. The simplest systematic plan for making predictions objectively from a single independent

¹ The study of time series is not included in this chapter. Brief reference to the general procedure of forecasting from such data was made on pages 235 and 238-39. The reader who is interested in prediction from historical statistics will find treatments in a number of texts. A good reference is Chaddock, R. E. *Principles and Methods of Statistics*. Boston: Houghton Mifflin Company, 1925. Chapter XIII.

variable is to transform the prognostic measures to the scale of the dependent variable and to use the transformed measures as the predictions. If \bar{X}_0 designates the transformed measures or the predictions, the formula is derived as follows:

$$\frac{\bar{X}_0 - M_0}{\sigma_0} = \frac{X_1 - M_1}{\sigma_1}$$

$$\bar{X}_0 = \frac{\sigma_0}{\sigma_1} (X_1 - M_1) + M_0$$

If the predictions are to be made from two or more prognostic measures (independent variables) they may be combined into a weighted sum. This derived variable is then used as X_1 in the above formula.

If X_0 is taken as the criterion, the error of prediction or estimate is represented by $X_0 - \bar{X}_0$. When the above procedure is employed, the magnitude of the errors of estimate for a typical population will be somewhat larger than when the regression equation is used as the formula of prediction. If the relationship between the two variables X_0 and X_1 is linear, the regression equation ¹ is

$$\bar{X}_0 = r_{01} \frac{\sigma_0}{\sigma_1} X_1 + M_0 - r_{01} \frac{\sigma_0}{\sigma_1} M_1$$

If the predictions are to be made from two or more independent variables, X_1, X_2, X_3 , etc., the multiple regression equation is

$$\bar{X}_0 = b_{01.23 \dots n} X_1 + b_{02.134 \dots n} X_2 \\ + \dots + b_{0n.123 \dots (n-1)} X_n + C$$

As in the case of predictions made from a single independent variable, the constants in the right-hand member of the equation are to be defined so that $\Sigma(X_0 - \bar{X}_0)^2$ will be a minimum.

¹The basis of the derivation of the regression equation is the requirement that $\Sigma(X_0 - \bar{X}_0)^2$ be a minimum. For the derivation, see Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, p. 159.

This condition is satisfied ¹ if

$$b_{01.23\dots n} = r_{01.234\dots n} \frac{\sigma_{0.234\dots n}}{\sigma_{1.234\dots n}}$$

$$b_{02.134\dots n} = r_{02.134\dots n} \frac{\sigma_{0.134\dots n}}{\sigma_{2.134\dots n}}$$

.....

$$b_{0n.123\dots (n-1)} = r_{0n.123\dots (n-1)} \frac{\sigma_{0.123\dots (n-1)}}{\sigma_{n.123\dots (n-1)}}$$

$$C = M_0 - b_{01.234\dots n}M_1 - b_{02.134\dots n}M_2 \\ - \dots - b_{0n.123\dots (n-1)}M_n$$

If the variables are expressed in terms of deviation measures, the constant term is zero and we have

$$\bar{x}_0 = b_{01.23\dots n}x_1 + b_{02.134\dots n}x_2 + \dots + b_{0n.123\dots (n-1)}x_n$$

If the variables are expressed in terms of standard deviation measures, β (beta) is used as the designation of the regression coefficients.

$$\bar{z}_0 = \beta_{01.23\dots n}z_1 + \beta_{02.134\dots n}z_2 + \dots + \beta_{0n.123\dots (n-1)}z_n$$

This is referred to as the *standard score regression equation*. The relation between the two types of regression coefficients is

$$b_{01.23\dots n} = \beta_{01.23\dots n} \frac{\sigma_0}{\sigma_1}$$

Hence the general form of the regression equation may be written

$$\bar{X}_0 = \beta_{01.23\dots n} \frac{\sigma_0}{\sigma_1} X_1 + \beta_{02.134\dots n} \frac{\sigma_0}{\sigma_2} X_2 \\ + \dots + \beta_{0n.123\dots (n-1)} \frac{\sigma_0}{\sigma_n} X_n + C$$

When the relationship between the two variables is not linear, an equation of prediction may be derived by the methods of

¹ For the proof of this statement see Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, pp. 283 f.

curve-fitting.¹ This equation may be any one of a number of types such as

$$\bar{X}_0 = AX_1^2 + C$$

$$\bar{X}_0 = A \log X_1 + C$$

$$\bar{X}_0 = \frac{1}{AX_1 + C}$$

Toops has proposed a "generalized regression" equation for which he claims superiority as a formula of prediction. The general form of Toops' equation² is

$$\bar{X}_0 = [1 + f_1(X_1)] \cdot [1 + f_2(X_2)] \cdot [1 + f_3(X_3)] \dots$$

The determination of the constants in the selected type of equation may be accomplished by the method of averages which imposes the requirement that the algebraic sum of the differences, $x_0 - \bar{x}_0$, or *residuals* equals or approaches zero, or by the method of least squares which imposes the requirement that the sum of the squares of the residuals be a minimum. Sometimes, more complex methods are used.

It should be noted that a prediction formula is derived from measures of one population and then applied to another. In other words, the population, from which the data for deriving the equation were obtained, is considered a representative sample of a larger population. Obviously, the accuracy of the predictions will be conditioned by the representativeness of the sample. Frequently, the requirement of representativeness can be only approximated. For example, in deriving a formula for predicting the college success of high school graduates, one is limited in securing data to high school graduates who enter college and remain long enough to secure marks. Since this group is somewhat selected, a sample of college freshmen will

¹ See Holzinger, *op. cit.*, pp. 317 f.

Ezekiel, Mordecai. *Methods of Correlation Analysis*. New York: John Wiley and Sons, Inc., 1930, Chapter VI.

² Toops, H. A. "Empirical Psychology and the 'Generalized Regression' Equation," *Ohio College Association Bulletin*, No. 81. Columbus, Ohio: Ohio State University, 1929, p. 1005.

not be entirely representative of the population of high school graduates.

Prediction by graphical methods. When predictions are to be made from a single independent variable, a graphical procedure may be employed. One advantage of this method is that it is not necessary to assume a linear relationship between the variables. The data are tabulated in a correlation table with the scale of the independent variable in the horizontal position. Then the mean of each column is calculated. These means are plotted as ordinates and the corresponding mid-points of the intervals of the horizontal scale, as abscissas. Through the points thus located, the best fitting curve is drawn. In drawing the curve, it should be recognized that the means at the extremes are based upon few cases and hence probably are not highly reliable. Hence, the curve may not fit the extreme points as closely as those for means based upon the larger number of cases.

In Figure 4 the means of the columns are given at the bottom and the line labeled *A* connects the points located by using them as ordinates. Since the relationship is apparently linear, the regression line labeled *B* in the figure probably represents the best fitting curve. Except at the extremes where the means are not dependable because the number of cases is small, prediction from line *A* will be approximately the same as those made from the regression equation ¹ represented by line *B*. For example, a child whose Otis Score (X_1) is 35 would be likely to secure a New Stanford Reading Score of 76 or 77 (\bar{X}_0). The value obtained from the regression equation is 76.38.

Prediction from the curve of relation versus prediction from the regression equation. The use of a regression equation as a formula of prediction gives the procedure the appearance of a high degree of accuracy. This is unfortunate. Even when the regression equation has been derived from a sample representative of the population within which predictions are to be made, the error of estimate, $X_0 - \bar{X}_0$, will approach zero in only a

¹ The regression equation is $\bar{X}_0 = .98X_1 + 42.08$. See page 324.

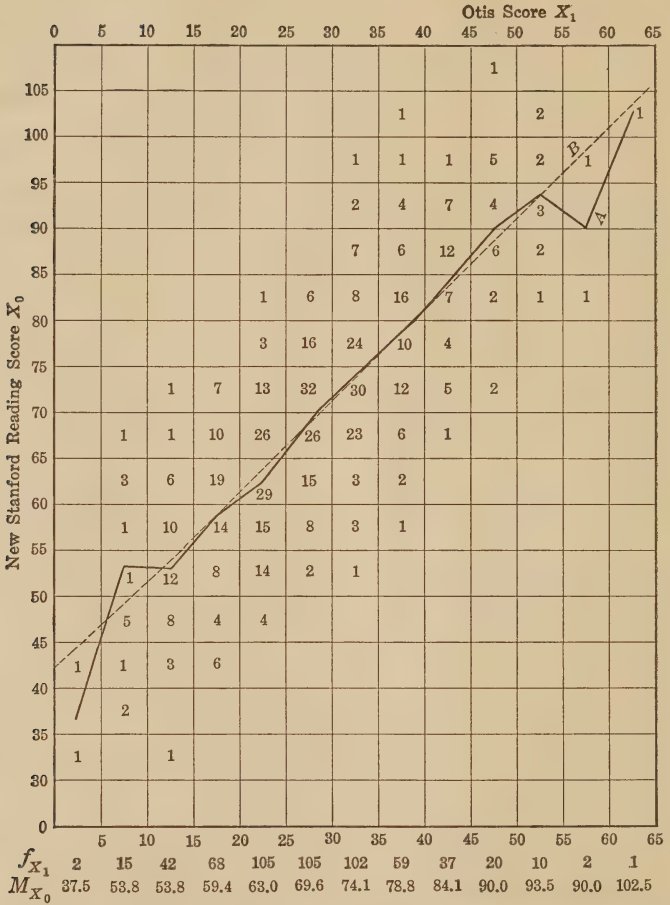


FIG. 4. Curves for predicting New Stanford Reading Test Scores from scores on the Otis Self-Administering Test of Mental Ability.

few cases unless r_{01} is very near to 1.00. In a typical prediction situation a number of the errors of estimate will be relatively large. This fact is apparent when the regression line is drawn on a correlation chart, because the error of prediction will be zero in the case of only the pairs of data which form coördinates of points on the line. Since the regression line passes through relatively few points, it follows that in general the predictions involve errors. The obviousness of the errors of estimate may be cited as an advantage of the graphical method, especially when predictions are being made by persons not familiar with regression equations and errors of estimate. Another advantage is that linearity of regression is not a requirement. A line can be drawn to fit any curvilinear relationship that is apparent. A third advantage is that after the line of relationship has been drawn, predictions may be made more quickly than when values are substituted in an equation.¹ Hence, when only one independent variable is involved, the graphical method is to be preferred as a practical procedure.²

The graphical method is also applicable to time series such as enrollment statistics over a period of years. As a means of making the general trend more apparent, the distribution may be smoothed by the method of moving averages. The simplest moving average is formed by taking the mean of the data for three successive years as the "smoothed" enrollment for the middle year of the sequence. The graphical representa-

¹ Both Hull and Segel have described a method for obtaining predictions automatically from electric tabulating and accounting machines. Such a method results in a great saving of time when a large number of predictions are to be made from a multiple regression equation.

Hull, C. L. "An Automatic Machine for Making Multiple Aptitude Forecasts," *Journal of Educational Psychology*, 16: 593-98, December, 1925.

Segel, David. "The Automatic Prediction of Scholastic Success by Using the Multiple Regression Equation Technique with Electric Tabulating and Accounting Machines," *Journal of Educational Psychology*, 22: 139-44, February, 1931.

² Griffin has shown how predictions from a multiple regression equation may be accomplished by graphical methods.

Griffin, Harold D. "Constructing a Prediction Chart (Charting Linear Regression Equations)," *Journal of Applied Psychology*, 16: 406-12, August, 1932.

tion of the "smoothed" enrollments affords a means of forecasting future enrollments.¹

Technique of deriving the regression equation when there are two or more independent variables. For two independent variables, the regression equation may be written as follows:

$$\bar{X}_0 = \frac{\sigma_0(r_{01} - r_{02}r_{12})}{\sigma_1(1 - r_{12}^2)} X_1 + \frac{\sigma_0(r_{02} - r_{01}r_{12})}{\sigma_2(1 - r_{12}^2)} X_2 + C$$

$$C = M_0 - \frac{\sigma_0(r_{01} - r_{02}r_{12})}{\sigma_1(1 - r_{12}^2)} M_1 - \frac{\sigma_0(r_{02} - r_{01}r_{12})}{\sigma_2(1 - r_{12}^2)} M_2$$

When written in this form, the plan of computation is apparent. It is only necessary to know M_0 , M_1 , M_2 , σ_0 , σ_1 , σ_2 , r_{01} , r_{12} , and r_{02} .

The regression equation for any number of independent variables may be derived from the general form given on pages 324-25. The regression coefficients (b 's) consist of the product of a coefficient of partial correlation² and the ratio of two partial standard deviations.

$$b_{01.234 \dots n} = r_{01.234 \dots n} \frac{\sigma_{0.234 \dots n}}{\sigma_{1.234 \dots n}}$$

A coefficient of partial correlation of any order can be expressed in terms of coefficients of lower order. For example, a first order partial may be expressed as follows:

$$r_{01.2} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1 - r_{02}^2} \sqrt{1 - r_{12}^2}}$$

In the case of a partial of $(n - 1)$ th order for $n + 1$ variables³ the relationship is

$$r_{01.23 \dots n} = \frac{r_{01.23 \dots (n-1)} - r_{0n.23 \dots (n-1)}r_{1n.23 \dots (n-1)}}{\sqrt{1 - r_{0n.23 \dots (n-1)}^2} \sqrt{1 - r_{1n.23 \dots (n-1)}^2}}$$

This general relationship makes it possible to build up a partial

¹ For further treatment of time series, see Chaddock, *op. cit.*, pp. 306 f.

² See pp. 377 f.

³ In this case the variables are $x_0, x_1, x_2, \dots, x_n$. If the variables are $x_1, x_2, x_3, \dots, x_n$, the order is $(n - 2)$ for n variables.

correlation coefficient of any order from zero order coefficients. The general formula for obtaining the partial standard deviations is

$$\sigma_{0.123\dots n} = \sigma_0 \sqrt{1 - r_{01}^2} \sqrt{1 - r_{02.1}^2} \sqrt{1 - r_{03.12}^2} \cdots \sqrt{1 - r_{0n.123\dots(n-1)}^2}$$

These general formulae for regression coefficients, coefficients of partial correlation, and partial standard deviations provide the basis for determining the regression equation for any number of variables in terms of their means, standard deviations, and intercorrelations. It is apparent, however, that the arithmetical work indicated by these formulae will be very laborious when the number of variables is greater than three (two independent variables). Some economies may be effected by systematizing the procedure,¹ but the maximum economy is attained by a method known as "partial regression."² This method is based upon the following relationships known as "normal equations":³

¹ For example, Garrett gives in detail a plan of the calculations for three, four, and five variables.

Garrett, H. E. *Statistics in Psychology and Education*. New York: Longmans, Green and Company, 1926, pp. 228 f.

² The reader who wishes to make a study of this topic will find a brief historical account in Griffin, H. D. "Partial Correlation versus Partial Regression for Obtaining Multiple Regression Constants," *Journal of Educational Psychology*, 22: 35-44, January, 1931.

This article gives references that may be consulted for further study. Some more recent references are:

Bakst, A. "A Modification of the Multiple Correlation and Regression Coefficients by the Tolley and Ezekiel Method," *Journal of Educational Psychology*, 22: 629-35, November, 1931.

Horst, A. "A General Method for Evaluating Multiple Regression Constants," *Journal of American Statistical Association*, 27: 270-78, September, 1932.

Peters, C. C., and Wykes, E. C. "Simplified Methods for Computing Regression Coefficients and Partial and Multiple Correlations," *Journal of Educational Research*, 23: 383-93, May, 1931.

³ Tolley, H. R., and Ezekiel, M. J. B. "A Method of Handling Multiple Correlation Problems," *Journal of American Statistical Association*, 18: 993-1003, December, 1923.

Garrett, H. E. "A Modification of Tolley and Ezekiel's Method of Handling Multiple Correlation Problems," *Journal of Educational Psychology*, 19: 45-49, January, 1928.

The equations derived by Tolley and Ezekiel are not in the form given here, but Garrett shows how they may be transformed by certain substitutions. Note that his symbolism is different from that employed here.

marks received. Such prediction has been called differential.¹ The predictions \bar{X}_D are subject to an error of estimate σ_{D-1} . If it is assumed that the differences X_D form a normal distribution with the mean at zero, any predictions may be interpreted in terms of the probability that the actual difference in average standing in the two fields will have the same sign as the predicted difference.²

In order to secure measures of the dependent variable, difference in average standing in the two fields, it is necessary that a representative population attempt simultaneously to achieve in both of the fields. This requirement creates a serious practical difficulty, especially in the case of fields of vocational activity. As a means of avoiding this difficulty, the same prognostic measures may be secured for a representative population in each of the two fields of endeavor. If the regression equations derived from these sets of data are expressed in standard score form, the predictions for a given individual will indicate the probable relative degree of success in the two fields.³

It should be noted that our present prognostic measuring instruments have been developed for ordinary prediction, and hence they may not be highly efficient for differential prediction. As the latter type of prediction is studied, it may be that superior differential prognostic instruments will be identified. The practical value of differential prediction appears to justify research directed to this end.

B. MEASURES OF THE ACCURACY OF PREDICTIONS

Standard error of estimate—one independent variable. The accuracy of predictions made by a given formula may be determined by applying it within a representative population, ob-

¹ Segel, David. "Differential Prediction of Ability as Represented by College Subject Groups," *Journal of Educational Research*, 25: 14-26, 93-98, January and February, 1932.

² For the technique by means of which this may be accomplished, see page 106.

³ For an illustration of this type of differential prediction, see Waits, J. V. "The Differential Predictive Value of the Psychological Examination of the American Council on Education," *Journal of Experimental Education*, 1: 264-71, March, 1933.

taining the criterion measures and computing the errors of estimate. For example, a formula for predicting scholastic success of high school graduates when they enter college may be applied to a representative population of college entrants. At the end of the year the marks received may be secured and the errors of estimate calculated. An "average" of these errors will be a measure of the accuracy of the predictions.

This method is simple and reveals the total error of estimate due to all causes operating in the case of the population to which the application is made. It, however, requires time, usually a year or longer to secure the criterion measures. For this reason, the formula is usually tried out with respect to a theoretical, normally distributed population for which the means and standard deviations are the same as these statistics for the population from which the data for deriving the formula were obtained. If the measured status is the criterion, the error of estimate is represented by $X_0 - \bar{X}_0$. If the true measure of the status is the criterion, the error of estimate is represented ¹ by $X_\infty - \bar{X}_0$. Within a typical population, some of the errors of estimate will be positive, others will be negative. If the assumption is made that for the population being considered the errors of estimate form a normal distribution with the mean at zero, the "average" magnitude may be measured by the standard deviation of this distribution. This statistic is called the *standard error of estimate*. The probable error of estimate may be obtained by multiplying by the constant .6745.

Since we have two types of formulae for making predictions when the relationship is linear and two criteria with which they may be compared, there are four standard errors of estimate. The method of deriving the formulae for these standard errors of estimate may be illustrated by considering the case in which the predictions are made from the regression equation and X_0 is taken as the criterion. It is assumed that $\bar{M}_0 = M_0$ which is

¹ The symbol ∞ designates "infinity" and X_∞ designates the mean of an infinite number of measures of the same individual, none of which involves a systematic error.

equivalent to saying that neither X_0 nor X_1 in the theoretical population involves a systematic error or that their systematic errors are equivalent. It is also assumed that the regression of X_0 upon X_1 is linear. Since $M_0 = \bar{M}_0$, $X_0 - \bar{X}_0 = (X_0 - M_0) - (\bar{X}_0 - \bar{M}_0) = x_0 - \bar{x}_0$, we may deal with deviations instead of raw measures. The regression equation in terms of deviation measures is:

$$\bar{x}_0 = r_{01} \frac{\sigma_0}{\sigma_1} x_1$$

Hence the error of estimate, $x_0 - \bar{x}_0$ is equal to

$$x_0 - r_{01} \frac{\sigma_0}{\sigma_1} x_1$$

The standard deviation of the distribution of the errors of estimate, or the standard error of estimate,¹ is represented by the symbol $\sigma_{0.1}$.

$$\begin{aligned} \sigma_{0.1}^2 &= \frac{\Sigma(x_0 - \bar{x}_0)^2}{N} \\ &= \frac{1}{N} \Sigma(x_0 - r_{01} \frac{\sigma_0}{\sigma_1} x_1)^2 \\ &= \frac{\Sigma x_0^2}{N} - 2r_{01} \frac{\sigma_0}{\sigma_1} \frac{\Sigma x_0 x_1}{N} + r_{01}^2 \frac{\sigma_0^2}{\sigma_1^2} \frac{\Sigma x_1^2}{N} \\ &= \sigma_0^2 - 2r_{01} \sigma_0^2 \frac{\Sigma x_0 x_1}{N \sigma_0 \sigma_1} + r_{01}^2 \sigma_0^2 \\ &= \sigma_0^2 - r_{01}^2 \sigma_0^2 \\ \sigma_{0.1} &= \sigma_0 \sqrt{1 - r_{01}^2} \end{aligned}$$

The expression $\sqrt{1 - r^2}$, which appears in a number of equations, has been called the *coefficient of alienation*² and is designated by k . Hence we may write $\sigma_{0.1} = \sigma_0 k_{01}$.

¹ The ordinary formula for the standard deviation is $\sigma = \sqrt{\frac{\Sigma x^2}{N}}$ where x represents the deviations from the mean. In the equation given here, x is replaced by the expression for the error of estimate and both sides are squared.

² Kelley, T. L. "Principles Underlying the Classification of Men," *Journal of Applied Psychology*, 3: 50-67, March, 1919.

The standard error of estimate ($\sigma_{0.1}$) gives the magnitude of the error that will be exceeded in approximately one-third of the predictions. The probable error of estimate ($PE_{0.1} = .6745\sigma_0\sqrt{1-r_{01}^2}$) is more easily interpreted because it gives the magnitude of the error that will be exceeded in one-half of the predictions.

If the estimates are compared with true measures of the dependent variable, the error of estimate is $X_\infty - \bar{X}_0$. The corresponding standard error of estimate is given by the equation

$$\sigma_{\infty.1} = \sigma_\infty \sqrt{1 - r_{\infty 1}^2}$$

$$\sigma_\infty = \sigma_0 \sqrt{r_{00}} \quad (\text{See page 151})$$

$$r_{\infty 1} = \frac{r_{01}}{\sqrt{r_{00}}} \quad (\text{See page 148})$$

Substituting these values in the above equation and simplifying, we have

$$\sigma_{\infty.1} = \sigma_0 \sqrt{r_{00} - r_{01}^2}$$

We may also write

$$PE_{\infty.1} = .6745\sigma_0 \sqrt{r_{00} - r_{01}^2}$$

If the predictions are made from the formula

$$\bar{X}_0 = \frac{\sigma_0}{\sigma_1}(X_1 - M_1) + M_0$$

the corresponding standard errors of estimate are given by the expressions $\sigma_0\sqrt{2-2r_{01}}$ and $\sigma_0\sqrt{1+r_{00}-2r_{01}}$.

Thus, we have four standard errors of estimate, two for predictions made from the linear regression equation and two for the use of transformed measures of the independent variable as predictions. Each is correct but the last two ¹ are seldom used because we are usually concerned with predictions made from the linear regression equation.

Accuracy of prediction when there are two or more independent variables. When the predictions of X_0 are obtained from two or more independent variables by means of a multiple

¹ No symbols have been proposed for designating them.

regression equation, the coefficient of correlation between these predictions (\bar{X}_0) and X_0 is given by the formula ¹

$$R_{0.123\dots n} = \sqrt{1 - [(1 - r_{01}^2)(1 - r_{02.1}^2)(1 - r_{03.12}^2)\dots(1 - r_{0n.12\dots(n-1)}^2)]}$$

The equation for the standard error of estimate is

$$\begin{aligned}\sigma_{0.123\dots n} &= \sigma_0 \sqrt{1 - R_{0.123\dots n}^2} \\ &= \sigma_0 \sqrt{(1 - r_{01}^2)(1 - r_{02.1}^2)\dots(1 - r_{0n.12\dots(n-1)}^2)}\end{aligned}$$

We may also write

$$\sigma_{\infty.123\dots n} = \sigma_0 \sqrt{r_{00} - R_{0.123\dots n}^2}$$

when \bar{X}_0 is taken as the estimate of $_{\infty}X_0$.

If the transformed weighted sum of the independent variables is used as the prediction, the corresponding standard errors of estimate would be obtained by replacing the coefficient of multiple correlation by the coefficient of correlation between X_0 and the weighted sum.

Procedures for interpreting a standard or probable error of estimate for a group of predictions. The fact that the standard error of estimate has been found to be a certain magnitude such as 3.7 or 18.3 is not very meaningful. One plan of interpretation is to determine the per cent of errors of estimate that are not greater than a certain amount. This plan is useful when the predictions are made in terms of a small number of categories such as school marks. For example, the interpretation may be made in terms of the per cent of marks correctly predicted, the per cent of predictions in error by one step of the scale, and so on. A second plan of interpretation is to compare the obtained standard error of estimate with the standard error of estimate for a criterion prediction. Two such bases of comparison have been employed—(1) chance or random prediction and (2) the mean of the dependent variable taken as the prediction for all members of the population. The result of these comparisons has been called “efficiency of prediction.” In both cases the

¹ It should be noted that $R_{0.123\dots n}$ is merely $r_{X_0\bar{X}_0}$.

interpretation is related to the coefficient of correlation between X_0 and X_1 . This relationship makes possible the determination of the accuracy or efficiency of the predictions from merely the coefficient of correlation.

Per cent of errors of estimate that are within a specified limit. When the predictions have been made from the regression equation and X_0 is taken as the criterion, the errors of estimate for a typical population are assumed to form a normal distribution whose standard deviation (standard error of estimate) is given by the formula

$$\sigma_{0.1} = \sigma_0 \sqrt{1 - r_{01}^2}$$

The problem is to determine the per cent of this distribution that lies within a specified distance from the mean which is zero. This distance is commonly described in terms of the scale on which X_0 and consequently the predictions are expressed. If σ_0 is taken as a unit of measurement, the specified limits may be expressed as $\pm m\sigma_0$. Solving the formula for the standard error of estimate, we have

$$\sigma_0 = \frac{\sigma_{0.1}}{\sqrt{1 - r_{01}^2}}$$

Multiplying both sides by m , we obtain

$$m\sigma_0 = \frac{m}{\sqrt{1 - r_{01}^2}} \sigma_{0.1} = \frac{m}{k_{01}} \sigma_{0.1}$$

The specified limit $m\sigma_0$ is thus equivalent to $\frac{m}{k_{01}} \sigma_{0.1}$. The per cent of the predictions whose errors of estimate fall within $\pm m\sigma_0$ may be found by dividing m by $\sqrt{1 - r_{01}^2}$ or k_{01} , locating this quotient as a deviation in a table ¹ of the areas under the normal probability curve corresponding to deviations from the mean and multiplying the corresponding area by 2. The per cents for various values of m and r_{01} are given in Table XI.

¹ For example, Holzinger, K. J. *Statistical Tables for Students in Education and Psychology*. Chicago: University of Chicago Press, 1925—Table XI. For description of this table and its uses see pages 80–81 of this book.

TABLE XI. SHOWING FOR VARIOUS VALUES OF r_{01} , THE PER CENT OF PREDICTIONS WHOSE ERROR IS NOT GREATER THAN THE AMOUNT INDICATED

r_{01}	PER CENT OF PREDICTIONS WHOSE ERROR OF ESTIMATE IS NOT GREATER THAN THE AMOUNT INDICATED											
	.1 σ_0	.2 σ_0	.3 σ_0	.4 σ_0	.5 σ_0	.6 σ_0	.8 σ_0	1.0 σ_0	1.25 σ_0	1.5 σ_0	2.0 σ_0	2.5 σ_0
.00	08	16	24	31	38	45	58	68	79	87	95	99
.10	08	16	24	31	38	45	58	68	79	87	96	99
.20	08	16	24	32	39	46	59	69	80	87	96	99
.30	08	17	25	33	40	47	60	71	81	88	96	99
.40	09	17	26	34	42	49	62	72	83	90	97	99
.50	09	18	27	36	44	51	64	75	85	92	98	100
.55	10	19	28	37	45	53	66	77	87	93	98	100
.60	10	20	29	38	47	55	68	79	88	94	99	100
.65	11	21	31	40	49	57	71	81	90	95	99	100
.70	11	22	33	42	52	60	74	84	92	96	99	100
.75	12	24	35	45	55	64	77	87	94	98	100	100
.80	13	26	38	50	59	68	82	90	96	99	100	100
.85	15	30	43	55	66	75	87	94	98	100	100	100
.90	18	35	51	64	75	83	93	98	100	100	100	100
.95	25	48	66	80	89	95	99	100	100	100	100	100
1.00	100	100	100	100	100	100	100	100	100	100	100	100

The application of this technique may be illustrated by assuming a five-point marking system—A, B, C, D, and E—in which the distribution of the marks conforms to the normal probability curve. If the total range is taken as 5σ and each mark represents a range of 1.00σ , the limit of the error of estimate for marks predicted correctly would be $.50\sigma$. For those in error by not more than one step, the limit would be 1.50σ . If the coefficient of correlation between the basis of prediction and the marks to be predicted is $.60$, $k_{01} = .80$. Hence, the magnitude of errors for marks predicted correctly would be

$$\frac{.50}{.80}\sigma_{0.1} = .625\sigma_{0.1}$$

When this distance is measured in both directions from the mean, 47 per cent of the area under the normal probability curve is marked off. Hence, when $r_{01} = .60$ and the assumptions are satisfied, 47 per cent of the marks will be predicted correctly.

By a similar procedure it is found that 94 per cent will be predicted correctly or be in error by only one step.

The values given in Table XI are for X_0 as the criterion. The per cent of errors of estimate that do not exceed a specified magnitude may be determined also for X_∞ as the criterion. The only change in the procedure is that m is divided by $\sqrt{r_{00} - r_{01}^2}$ instead of by $\sqrt{1 - r_{01}^2}$. If the reliability of the criterion is .75 and $r_{01} = .60$, 57 per cent of the predictions will be within $.50\sigma_0$ of the true criterion. The per cents corresponding to various values of r_{00} , r_{01} , and m cannot be conveniently given because a separate table would be required for each value of r_{00} .

Efficiency of prediction when X_0 is the criterion. The second method of interpreting the standard error of estimate is to compare it with the standard error of estimate for chance or random predictions, or for the mean of the dependent variable (M_0) taken as the prediction for all members of the population. When M_0 is used as the prediction ¹ the standard error of estimate is σ_0 . Using this as a norm, the reduction or improvement in the standard error of the estimate when the predictions are made from the regression equation is given by the expression $\sigma_0 - \sigma_0\sqrt{1 - r_{01}^2}$. The per cent of improvement may be found by dividing this by σ_0 .

$$\frac{\sigma_0 - \sigma_0\sqrt{1 - r_{01}^2}}{\sigma_0} = 1 - \sqrt{1 - r_{01}^2}$$

The expression ² $(1 - \sqrt{1 - r_{01}^2})$ has been called "efficiency of prediction" or "predictive index" and E , I_p , and PI have been used as designations for it. As a means of avoiding confusion with the expression developed in the following paragraph, E_M is suggested as a symbol.

¹ The use of M_0 as the prediction for all members of the population corresponds to $r_{01} = 0$. This is obvious from the regression equation

$$\bar{X}_0 = r_{01}\frac{\sigma_0}{\sigma_1}x_1 + M_0 - r_{01}\frac{\sigma_0}{\sigma_1}M_1$$

If $r_{01} = 0$, the two terms involving it become zero and $\bar{X}_0 = M_0$.

² The factor, 100, is sometimes inserted as a multiplier to enable us to write the values obtained as per cents without a decimal point.

In order to serve as a convenient criterion, chance or random predictions must have the same mean and the same standard deviation as X_0 . For example, if letter grades of A, B, C, D, and E are being predicted, this requirement means that the number of A's predicted is equal to the number of A's actually received, the number of B's predicted is equal to the number of B's actually received, and so on. Estimates of the future status of a group of students by a teacher who is acquainted with them will probably not be random predictions¹ but will be correlated with X_0 . If the random predictions are represented by X_g , the error of estimate will be $X_0 - X_g$.

$$\begin{aligned}\sigma_{0.g} &= \sqrt{\frac{\Sigma(X_0 - X_g)^2}{N}} \\ &= \sqrt{\sigma_0^2 + \sigma_g^2}\end{aligned}$$

Since σ_g is assumed to be equal to σ_0 ,

$$\sigma_{0.g} = \sqrt{2}\sigma_0^2$$

The reduction in the standard error of estimate is

$$\sqrt{2}\sigma_0^2 - \sigma_0\sqrt{1 - r_{01}^2}$$

The per cent of improvement is obtained by dividing this expression by $\sqrt{2}\sigma_0^2$.

$$\frac{\sqrt{2}\sigma_0^2 - \sigma_0\sqrt{1 - r_{01}^2}}{\sqrt{2}\sigma_0^2} = 1 - \sqrt{\frac{1 - r_{01}^2}{2}}$$

¹ Kaulfers, W. V. "A Guessing Experiment in Foreign Language Prognosis," *School and Society*, 32: 535-38, October 18, 1930.

In this study a number of teachers were instructed to guess the final grades of their students at the beginning of the third day of the semester. The correlations of the "guesses" with the grades actually received range from .05 ($N = 23$) to .73 ($N = 75$), and eight of the seventeen coefficients are above .50. It is, of course, not known to what extent the estimates made influenced the final marks, but the resulting correlations indicate very clearly that the predictions made on the basis of only very limited acquaintance with pupils are likely to be far from "pure" guesses. In an unreported study by the senior author, estimates of final grades in chemistry were made by a graduate student, with no teaching experience, on the basis of very limited casual observation of the students in the laboratory. The correlation of the estimates with final marks was .28 ($N = 94$). When estimates were determined by arranging the names of the students in alphabetical order and assigning A's to the first ones in the list, B's to those next, and so on, the correlation between these estimates and the final marks was -.037 in one case and -.088 in a second.

The expression $1 - \sqrt{\frac{1 - r_{01}^2}{2}}$ is the per cent of improvement over "pure guesses" or chance predictions and may be designated by E_g .

If the predictions are made by means of the formula

$$\bar{X}_0 = \frac{\sigma_0}{\sigma_1}(X_1 - M_1) + M_0$$

the standard error of estimate is $\sigma_0 \sqrt{2 - 2r_{01}}$, and the corresponding measures of the efficiency of prediction are $1 - \sqrt{2 - 2r_{01}}$ and $1 - \sqrt{1 - r_{01}}$. These measures are new and P_M and P_g are suggested as symbols.

Thus we have two formulae for the "efficiency of prediction" when estimates are made from the linear regression equation, one giving the per cent of improvement over using the mean as the prediction for all members of the population and the other giving the per cent of improvement over "pure guesses." We also have two corresponding measures of the efficiency of prediction when the estimates are made from

$$\bar{X}_0 = \frac{\sigma_0}{\sigma_1}(X_1 - M_1) + M_0$$

For $r_{01} = .60$ these measures of efficiency of prediction are .20, .43, .11, and .37. All of these measures are correct but each must be properly interpreted. It is unfortunate that E_M , which is commonly used, has been interpreted as the "improvement over chance in prediction."¹ The result has been a false im-

¹ For example see Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, p. 166. The symbol I_p is used instead of E_M .

The only comment on this erroneous interpretation that has come to the attention of the present writers is in a recent article.

Douglass, H. R. "Some Observations and Data on Certain Methods of Measuring the Predictive Significance of the Pearson Product-Moment Coefficient of Correlation," *Journal of Educational Psychology*, 25: 225-31, March, 1934.

A few persons have correctly understood the formula $E_M = 1 - \sqrt{1 - r_{01}^2}$. For example, see: Bailor, E. M. "Content and Form in Tests of Intelligence," *Teachers College, Columbia University Contributions to Education*, No. 162. New York: Bureau of Publications, Teachers College, Columbia University, 1924, p. 25.

In view of the fact that Walker in *Studies in the History of Statistical Method*,

pression in regard to the predictive value of prognostic measures. Table XII gives both E_M and E_g for various values of r_{01} . These measures of predictive efficiency are related as follows:

$$E_g = .7071E_M + .2929$$

$$E_M = 1.4142E_g - .4142$$

It is, of course, permissible to use any measure of predictive efficiency provided it is correctly interpreted, but the present writers are of the opinion that our thinking relative to the predictive value of prognostic measures will be facilitated by employing

$$E_g = 1 - \sqrt{\frac{1 - r_{01}^2}{2}}$$

TABLE XII. EFFICIENCY OF PREDICTION OF X_0 BY MEANS OF REGRESSION EQUATION, FOR VARIOUS VALUES OF r_{01} OR $R_{0.12 \dots n}$

r_{01} OR $R_{0.12 \dots n}$	E_M^*	E_g^\dagger	r_{01} OR $R_{0.12 \dots n}$	E_M	E_g
.00	0	29.3	.55	16	41
.05	0.1	29.4	.60	20	43
.10	0.5	29.6	.65	24	46
.15	1.1	30.1	.70	29	50
.20	2.0	30.7	.75	34	53
.25	3.2	31.5	.80	40	58
.30	4.6	32.5	.85	47	63
.35	6	34	.866	50	65
.40	8	35	.90	56	69
.45	11	37	.95	69	78
.50	13	39	1.00	100	100

* Calculated by means of formula $E_M = 1 - \sqrt{1 - r_{01}^2}$.

† Calculated by means of formula $E_g = 1 - \sqrt{\frac{1 - r_{01}^2}{2}}$.

Values of P_M and P_g are given in Table XIII. It is interesting to note that the value of P_M is negative for r_{01} less than .50. This means that transformed values of X_1 are less efficient as predictions than M_0 used as the prediction for all members of

p. 186, credits Bailor with originating the term "predictive index" and presumably the formula, it is strange that the basis of its derivation has been overlooked by so many who have used it.

the population. Comparison of Tables XII and XIII will reveal the superiority of predictions made from the regression equation.

Efficiency of prediction when \bar{X}_∞ is the criterion. If the measures of X_0 are fallible, i.e., involve variable errors of measurement, and the predictions are to be compared with true measures of the dependent variable, X_∞ is the criterion. If M_0 is taken as the prediction for each member of the population, the standard error of estimate, standard deviation of $X_\infty - M_0$, is σ_∞ which is equal to $\sigma_0 \sqrt{r_{00}}$. Hence, we have $\sigma_{\infty \cdot M} = \sigma_0 \sqrt{r_{00}}$. If random predictions X_g are compared with the true measures, the error of estimate will be $X_\infty - X_g$.

$$\begin{aligned}\sigma_{\infty \cdot g} &= \sqrt{\frac{\Sigma(X_\infty - X_g)^2}{N}} \\ &= \sqrt{\sigma_\infty^2 + \sigma_g^2}\end{aligned}$$

Since $\sigma_\infty = \sigma_0 \sqrt{r_{00}}$ and $\sigma_g = \sigma_0$

$$\begin{aligned}\sigma_{\infty \cdot g} &= \sqrt{\sigma_0^2 r_{00} + \sigma_0^2} \\ &= \sigma_0 \sqrt{1 + r_{00}}\end{aligned}$$

TABLE XIII. EFFICIENCY OF PREDICTION OF x_0 , WHERE $\frac{\sigma_0}{\sigma_1} x_1$ IS TAKEN AS EVIDENCE OF x_0 , FOR VARIOUS VALUES OF r_{01} .

r_{01}	P_M^*	P_g^\dagger	r_{01}	P_M	P_g
.00	-41	0	.55	5	33
.05	-38	2.5	.60	11	37
.10	-34	5.1	.65	16	41
.15	-30	7.8	.70	23	45
.20	-26	10.6	.75	29	50
.25	-22	13	.80	37	55
.30	-18	16	.85	45	61
.35	-14	19	.90	55	68
.40	-10	23	.95	68	78
.45	- 5	26	1.00	100	100
.50	0	29			

* Calculated by means of formula: $P_M = 1 - \sqrt{2 - 2r_{01}}$.

† Calculated by means of formula: $P_g = 1 - \sqrt{1 - r_{01}}$.

Comparing $\sigma_{\infty \cdot 1} = \sigma_0 \sqrt{r_{00} - r_{01}^2}$ with these criterion standard errors of estimate we obtain ¹

$$E_{\infty M} = 1 - \frac{\sqrt{r_{00} - r_{01}^2}}{\sqrt{r_{00}}}$$

$$E_{\infty g} = 1 - \frac{\sqrt{r_{00} - r_{01}^2}}{\sqrt{1 + r_{00}}}$$

From the standard error of estimate $\sigma_0 \sqrt{1 + r_{00} - 2r_{01}}$ two additional measures of predictive efficiency are obtained

$$P_{\infty M} = 1 - \frac{\sqrt{1 + r_{00} - 2r_{01}}}{\sqrt{r_{00}}}$$

$$P_{\infty g} = 1 - \frac{\sqrt{1 + r_{00} - 2r_{01}}}{\sqrt{1 + r_{00}}}$$

Table XIV gives $E_{\infty g}$ for various values of r_{00} and r_{01} or $R_{0.123 \dots n}$. The column for $r_{00} = 1.00$ gives the values for E_g . Comparison of a value in this column with the corresponding ones in preceding columns shows how much more efficiently we are able to predict true measures of the criterion than fallible measures of it.

For example, if the coefficient of correlation between scores on an aptitude test and the marks received in a course (r_{01}) is .60, the improvement over chance prediction of the marks is .43. If the coefficient of reliability of the marks (r_{00}) is .70, $E_{\infty g} = .55$. Hence the efficiency of predicting the fallible marks is .43 per cent better than pure guess, while the efficiency of predicting "true" marks is .55 per cent better than pure guess. It should be noted that except by chance, r_{01} , the correlation between the measures used in prediction and the fallible criterion scores, cannot be greater than the coefficient of reliability of the criterion scores.

¹ For a different form of the first formula, see Conrad, H. S., and Martin, G. B. "The Index of Forecasting Efficiency for the Case of a 'True' Criterion," *Journal of Experimental Education*, March, 1936.

TABLE XIV. EFFICIENCY OF PREDICTION OF \bar{X}_∞ BY MEANS OF REGRESSION EQUATIONS, FOR VARIOUS VALUES OF r_{01} OR $R_{0.12\dots n}$ AND r_{00} .

r_{01} OR $R_{0.12\dots n}$	r_{00}														
	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00
.00	52	49	47	44	42	40	39	37	36	35	33	32	31	30	29
.05	52	49	47	44	42	41	39	37	36	35	33	32	31	30	29
.10	53	50	47	45	43	41	39	38	36	35	34	33	32	31	30
.15	54	51	48	46	44	42	40	38	37	36	34	33	32	31	30
.20	55	52	49	47	45	43	41	39	38	36	35	34	33	32	31
.25	57	54	51	48	46	44	42	40	39	37	36	35	34	33	32
.30	60	56	53	50	48	46	44	42	40	39	37	36	35	34	33
.35		59	55	52	50	47	45	43	42	40	39	37	36	35	34
.40			59	55	52	50	48	46	44	42	40	39	38	36	35
.45				59	55	53	50	48	46	44	42	41	39	38	37
.50					59	56	53	51	49	47	45	43	42	40	39
.55						60	57	54	52	49	47	46	44	42	41
.60							61	58	55	53	51	49	47	45	43
.65								63	60	57	54	52	50	48	46
.70									65	61	59	56	54	51	50
.75										67	64	61	58	55	50
.80											70	66	63	60	53
.85												74	69	66	63
.90													78	73	69
.95														84	78
1.00															1.00

Calculated by means of formula

$$E_{\infty g} = 1 - \frac{\sqrt{r_{00} - r_{01}^2}}{\sqrt{1 + r_{00}}}$$

Figure 5 shows graphically the values of E_M and E_g , and of $E_{\infty g}$ for r_{00} equal to .50 and .70. This figure affords an excellent basis for arriving at a clear understanding of these measures of the efficiency of prediction.

Efficiency of prediction when a multiple regression equation is used. The preceding exposition of procedures for interpreting the standard error of estimate of predictions has been in terms of prediction from a single independent variable. If a multiple regression is used $R_{0.123\dots n}$ may be substituted for r_{01} . It should be noted, however, that the coefficient of multiple correlation tends to exaggerate the accuracy of the predictions

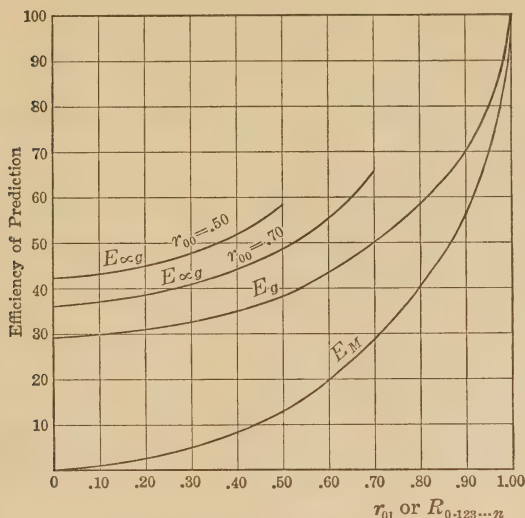


FIG. 5. Different measures of efficiency of prediction for values of r_{01} .

and hence for precise determinations the obtained $R_{0.123\dots n}$ should be corrected.¹

Accuracy of individual predictions. The preceding pages have dealt with the accuracy of predictions for a representative population as a group. The probable error of estimate may be used with individual predictions. The procedure is similar to that of interpreting the probable error of measurement with reference to individual scores.² When X_0 is the criterion and the prediction has been made from the linear regression equation, we may write

$$\text{Probable limits of status} = \bar{X}_0 \pm .6745\sigma_0\sqrt{1 - r_{01}^2}$$

¹ Ezekiel, M. *Methods of Correlation Analysis*. New York: John Wiley and Sons, 1930, p. 177. Ezekiel gives the following correction formula

$$\bar{R}_{0.123\dots n}^2 = 1 - (1 - R_{0.123\dots n}^2) \frac{n - 1}{n - m}$$

in which n is the number of sets of observations in the sample and m is the number of constants in the regression equation.

² See page 134.

When X_∞ is taken as the criterion, we have

$$\text{Probable limits of status} = \bar{X}_0 \pm .6745\sigma_0\sqrt{r_{00} - r_{01}^2}$$

Corresponding statements may be made when transformed values of X_1 are used as predictions.

The probable limits of the status affords a basis for determining the chances that the status actually attained will be above (or below) a specified position. Suppose, for example, it is desired to determine the chances that a student whose score on a prognostic test is X_1 will achieve a standing above a particular mark. An answer may be easily obtained from the correlation table provided the number of students having the specified score is sufficiently large. The theoretical answer may be derived in general form as follows: Suppose the specified position is represented by ¹ $M_0 + m\sigma_0$. The predicted status is given by the regression equation

$$\bar{X}_0 = M_0 + r_{01}\frac{\sigma_0}{\sigma_1}(X_1 - M_1)$$

The difference between the specified status and the predicted status is

$$M_0 + r_{01}\frac{\sigma_0}{\sigma_1}(X_1 - M_1) - (M_0 + m\sigma_0)$$

which simplifies to

$$\sigma_0\left(r_{01}\frac{X_1 - M_1}{\sigma_1} - m\right)$$

The predicted status, however, is subject to an error of estimate. This means that a student whose predicted status is below the specified status may attain a status above it, and that one whose predicted status is above may fall below. The probability of the attained status being above the specified status may be determined. The first step is to express the above difference in terms of $\sigma_{0.1}$, the standard deviation of the distribution of the errors of estimate. Since $\sigma_{0.1} = \sigma_0\sqrt{1 - r_{01}^2} = \sigma_0k_{01}$,

¹ The value of m may be either positive or negative.

$\sigma_0 = \frac{\sigma_{0.1}}{k_{01}}$. Substituting this value for σ_0 we have

$$\sigma_0 \left(r_{01} \frac{X_1 - M_1}{\sigma_1} - m \right) = \sigma_{0.1} \frac{r_{01} \frac{X_1 - M_1}{\sigma_1} - m}{k_{01}}$$

Given a value of X_1 and of m , the fraction is easily evaluated. The probability corresponding to it is determined by referring to a table ¹ giving areas under the normal probability curve for various distances from the mean. If the value of the expression is $+.6745\sigma_{0.1}$, the chances that the attained status will be above the specified status are 75 in 100. If the value is $-.6745\sigma_{0.1}$, the chances that the attained status will be above the specified status are 25 in 100. Conversely, the chances that the attained status will be below the specified status are in the first case 25 in 100 and in the second 75 in 100. If the value is $+1.175\sigma_{0.1}$, the chances are 88 in 100 that the attained status will be above the specified status.

Critical scores. In dealing with prognostic measures several workers have introduced the idea of a "critical score," i.e., the score below which the chances of a specified degree of success are less than a stated probability. A critical score is most conveniently obtained from the correlation table,² but a theoretical critical score may be determined from the relationship developed in the preceding paragraph. In this relationship the variables are m , X_1 , and the probability. Given m and the probability, we may determine the value of X_1 which will be the critical score. Let the deviation value corresponding to the given probability be represented by $m'\sigma_{0.1}$.

$$m'\sigma_{0.1} = \frac{r_{01} \frac{X_1 - M_1}{\sigma_1} - m}{k_{01}} \sigma_{0.1}$$

¹ See page 80 for reference to a convenient table.

² For an illustration see Torgerson, T. L., and Aamodt, Geneva P. "The Validity of Certain Prognostic Tests in Predicting Algebraic Ability," *Journal of Experimental Education*, 1: 277-79, March, 1933.

This illustration is for transformed values of X_1 used as predictions. A similar table could be constructed for predictions from the regression equation.

Solving for X_1 we have

$$X_1 = M_1 + \frac{m + m'k_{01}}{r_{01}} \sigma_1$$

Suppose the specified probability of failure is three in four. The corresponding value of m' is $-.6745$. Suppose also that m is $-.50$ which corresponds to defining the minimum limit of success as a mark of C in a normally distributed system of five marks A, B, C, D, and E. If $r_{01} = .60$, $X_1 = M_1 - 1.73\sigma_1$. Hence the theoretical critical score for the specified conditions is $1.73\sigma_1$ below the mean of the prognostic measures. Three out of every four students with such a score on the prognostic test may be expected to receive a D or E.

Coefficient of correlation as an index of predictive efficiency. One of the uses of the coefficient of correlation mentioned in Chapter IV, page 116, was that of determining the predictive values of prognostic measures. Since the formulae for efficiency of prediction involve r as a variable, the values given in Tables XII, XIII, and XIV are essentially interpretations of the corresponding coefficients of correlation when this statistical technique is employed to investigate the prognostic value of a given set of measures. For example, if $r_{01} = .60$, $E_M = .20$ and $E_g = .43$. Hence the coefficient of correlation may be referred to as an "index of predictive efficiency" or more simply as a "predictive index."

C. EFFICIENCY OF PREDICTIONS IN PRACTICE

Magnitude of the predictive index in practice. Our interest in prediction in education is mainly with respect to future success either in school or in a vocational activity. The magnitude of the obtained correlations, r_{01} or $R_{0.123 \dots n}$, vary but the situation may be illustrated by reviewing briefly reported studies relating to (1) predicting success in the first year of college, and (2) predicting teaching success.

The reader should note a significant difference between these two prediction situations. In the first, "success in the first year

of college" is commonly thought of as the mean of the marks the student receives and when defined in this way valid measures of the variable to be predicted are available. Teaching success, on the other hand, has no generally recognized measure and as will be pointed out later there is evidence that the ratings obtained are grossly lacking in validity.

1. *Predicting success in the first year of college.*¹ Since Thorndike² and others presented evidence to show that the traditional college entrance examination was not a satisfactory means of determining admission to college, the predictive value of several items of information has been investigated. The reported coefficients of correlation for marks received in high school vary rather widely, probably due largely to differences in the populations from which data were secured.³ The values of r range from about .30 to over .70 and the central tendency falls within the interval .50 and .55. If only the more comprehensive investigations are considered, it seems probable that for the typical college, the coefficient of correlation between mean high school standing and mean mark in the freshman year is in the neighborhood of .55. This estimate is supported by the investigation of Odell,⁴ who secured the records of 1677 students who graduated from Illinois high schools in 1924. The value of r was found to be .55.

The range of the coefficients of correlation for intelligence test scores and mean marks in first year of college is somewhat

¹ This brief summary does not adequately represent the amount of research relating to the prediction of scholastic success at the college level. The interested reader will find a number of references in the writings by Douglass, Odell, and Tyler. Kaulfers has reported a summary of the research relating to prognosis in foreign languages. Kaulfers, W. V. "Present Status of Prognosis in Foreign Language," *School Review*, 39: 585-96, October, 1931.

² Thorndike, E. L. "The Future of the College Entrance Examination Board," *Educational Review*, 31: 470-83, May, 1906.

³ For a brief summary, see Douglass, H. R. "The Relation of High School Preparation and Certain Other Factors to Academic Success at the University of Oregon," *University of Oregon Publication*, Vol. III, No. 1. Eugene: University of Oregon, 1931. 61 pp.

⁴ Odell, C. W. "Predicting the Scholastic Success of College Freshmen," *University of Illinois Bulletin*, Vol. 25, No. 2, *Bureau of Educational Research Bulletin*, No. 37. Urbana: University of Illinois, 1927. 54 pp.

less, extending from about .30 to slightly above .60. The central tendency is approximately .45. For some of the more appropriate tests a slightly higher coefficient will be obtained from a typical population. Age at graduation from high school, ratings by the principal, and scores yielded by various instruments for measuring personality traits¹ have been compared with achievement in college, but the predictive value appears to be relatively low and the inclusion of such items in a multiple regression equation does not result in very large improvements over the predictions made from high school record and intelligence test scores.

Odell² found a coefficient of multiple correlation of .58 for mean high school mark and the score made on the Otis Self-Administering Test of Mental Ability, Higher Examination as independent variables. This is only .03 larger than the simple coefficient of correlation between mean mark in first year of college and mean high school mark. Douglass³ reported corresponding coefficients of .56 and .63 using the percentile rank on the American Council on Education Psychological Test as the measure of intelligence. Wood⁴ using records of only 97 students found a multiple coefficient of .66 for mean high school mark, score on Thorndike Intelligence Examination for High School Graduates, and Score on New York Regents' Examination. A few other investigators⁵ have secured slightly higher coefficients. It appears, however, that considering the practical difficulties of obtaining additional information before the student enters college, the improvement of prediction does not justify the use of data other than high school record and

¹ For an illustration, see Tyler, H. T. "The Bearing of Certain Personality Factors Other Than Intelligence on Academic Success," *Teachers College, Columbia University Contributions to Education*, No. 468. New York: Bureau of Publications, Teachers College, Columbia University, 1931. 89 pp. This monograph gives a brief summary of the more important previous studies and a helpful bibliography.

² *Op. cit.*, p. 37.

³ *Op. cit.*, p. 48.

⁴ Wood, B. D. *Measurement in Higher Education*. Yonkers-on-Hudson: World Book Company, 1923, p. 87.

⁵ See Douglass, *op. cit.*, p. 50, for summary of findings.

score on an intelligence test designed to predict college success.

2. *Predicting teaching success.* Beginning with Meriam's pioneer investigation ¹ there have been numerous studies of the relation of marks in academic subjects, marks in professional subjects, general intelligence test scores, and measures of other traits to measures of teaching success.² The reported coefficients of correlation between general intelligence and teaching success range upward from approximately zero to about .45. The highest correlations are reported by Somers ³ who employed a composite of the Thurstone Cycle Omnibus Test and the Trabue Language Completion Test as the measure of intelligence. It appears likely that for a typical population of teachers the correlation between scores on an intelligence test and the ratings of supervisors is not much greater than .20.

The coefficients of correlation reported for measures of achievement in professional courses (education) and for scores made on professional tests also vary widely, but on the average are slightly higher. For general scholarship, mean mark in teaching subject, or other phases of academic training, the coefficients range upward from approximately zero to .60, reported by Somers who also found a correlation of .615 for ratings of personality made at the end of the freshman year of training. Hunt ⁴ has reported coefficients ranging from .30 to .50 for scores on an aptitude test, and Morris ⁵ found a

¹ Meriam, J. L. "Normal School Education and Efficiency in Teaching," *Teachers College, Columbia University Contributions to Education*, No. 1. New York: Bureau of Publications, Columbia University, 1906, pp. 51-115.

² For a tabular summary of the "best known" studies, see Barr, A. S., and Douglas, Lois. "The Pre-training Selection of Teachers," *Journal of Educational Research*, 28: 92-117, October, 1934.

The accompanying bibliography includes 172 references.

³ Somers, T. T. "Pedagogical Prognosis," *Teachers College, Columbia University Contributions to Education*, No. 140. New York: Bureau of Publications, Teachers College, Columbia University, 1923. 129 pp.

⁴ Hunt, Thelma. "Measuring Teaching Aptitude," *Educational Administration and Supervision*, 15: 334-42, May, 1929.

⁵ Morris, E. H. "Personal Traits and Success in Teaching," *Teachers College, Columbia University Contributions to Education*, No. 342. New York: Bureau of Publications, Teachers College, Columbia University, 1929. 75 pp.

coefficient of .510 for scores on a Trait Index Test and practice teaching marks.

The variation in the reported coefficients of correlation for measures of general intelligence and other traits is doubtless due in part to differences in the groups of teachers for which data were secured. In most cases the group studied does not appear to be very representative of the general population of teachers. Another and probably more significant explanation of the variation is the presence of variable errors in the obtained measures of teaching success. Since the measures commonly employed are subjective, variable errors of measurement are to be expected, and it is likely that the estimates involve also relatively large variable errors of validity. It seems reasonable to say that the true criterion of a teacher's success is the total growth of her pupils toward the objectives of the school. Measures of this criterion are difficult to secure, but studies ¹ of the correlation between ratings of teachers and certain measures of pupil achievement indicate that the validity of the ratings of teachers is so low as to make them practically worthless as measures of teaching success.² The ratings made by a supervisor are merely his estimates with reference to what he considers good teaching to be. It is generally assumed that a more valid measure of teaching success is secured by using a composite of several ratings.³ The reliability of the composite may be high ⁴

¹ Crabbs, L. M. "Measuring Efficiency in Supervision and Teaching," *Teachers College, Columbia University Contributions to Education*, No. 175. New York: Bureau of Publications, Columbia University, 1925. 98 pp.

Taylor, H. "The Influence of the Teacher on Relative Class Standing in Arithmetic Fundamentals and Reading Comprehension," *Twenty-Seventh Year-book of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1928, pp. 97-100.

² For a more extensive review of the evidence see Corey, S. M. "The Present State of Ignorance about Factors Effecting Teaching Success," *Educational Administration and Supervision*, 28: 481-90, October, 1932.

³ Boardman reports a coefficient of .33 between general intelligence scores and a composite of ratings. This is one of the highest that has come to the attention of the present writers.

Boardman, C. W. "Professional Tests as Measures of Teaching Efficiency in High School," *Teachers College, Columbia University Contributions to Education*, No. 327. New York: Bureau of Publications, Teachers College, Columbia University, 1928. 84 pp.

⁴ The reliability of Boardman's composite is given as .91.

but this condition does not reveal its validity. In view of our state of ignorance concerning what constitutes good teaching, we are not justified in assuming that the mean of any number of ratings approaches a valid measure of teaching success. The fact that we have no satisfactory means of measuring teaching success makes the value of any prediction formula problematical.¹

Dependability of an index of predictive efficiency. The derivation of the standard error of estimate and of the subsequent indices of predictive efficiency is based upon the assumption that the population to which the regression equation is applied as a formula of prediction is equivalent in all essential respects to the population from which the data for the derivation of the equation were obtained. The population to which the formula is applied in practice is likely not to be equivalent in all essential respects. For example, in securing data for deriving a formula to predict the success of high school graduates in their first year of college, one is limited to those graduates who enter college and remain long enough to permit measurement of their success. Such a group is not likely to be entirely representative of high school graduates in general. Furthermore, the courses pursued and the instructional conditions in college may vary from institution to institution and within the same institution from year to year. Lack of equivalence in any respect that affects the marks received will tend to make the index of predictive efficiency lacking in dependability.

Variable errors of validity in the criterion affect the dependability of the index of predictive efficiency. The probable presence of variable errors of validity in the measures of teaching success provides an explanation of the low index of predictive efficiency. In considering the effect of variable errors of validity in the criterion, it should be noted that their magnitude depends upon the label attached to the predictions. If they are considered merely estimates of what has been measured, there will,

¹ A good reference relative to this point is Haggerty, M. E. "The Crux of the Teaching Prognosis Problem," *School and Society*, 35: 545-49, April 23, 1932.

of course, be no variable errors of validity. We, however, usually wish to predict *actual* scholastic success, *actual* teaching success, or whatever the label specifies, and hence, variable errors of validity are usually possible. In the absence of valid measures of the criterion, the magnitude of the errors of estimate cannot be known and hence the dependability of the index of predictive efficiency cannot be determined.

In deriving the formula for the standard error of estimate, the requirement was made that $M_0 = \bar{M}_0$ so that $x_0 - \bar{x}_0$ could be substituted for $X_0 - \bar{X}_0$. (See page 335.) This requirement is equivalent to specifying that the measures from which the regression equation is derived and the measures used in making predictions do not involve a systematic error or that the systematic errors in the corresponding groups of measures are equivalent. Hence, the presence of a systematic error in an independent variable or in the dependent variable, either in the data from which the regression equation is derived or in the measures used in making predictions, will affect the errors of estimate unless the systematic errors are equivalent. When the errors of estimate are affected, the dependability of an index of predictive efficiency will also be affected. Edgerton ¹ has derived an equation for the standard error of estimate which includes the effect of systematic errors of measurement. The formula, however, is relatively complex and hence is seldom used.

The dependability of an index of the predictive efficiency is affected also by non-conformity of the data to assumptions made in the derivation of the regression equation. For example, the linear regression equation assumes linearity of relationship and hence is appropriate only when this assumption is satisfied. If it is not, the dependability of the index of predictive efficiency will be affected.²

¹ Edgerton, H. A. "Measuring the Validity of Predicted Scores," *Journal of Educational Psychology*, 21: 388-91, May, 1930.

² For a discussion of the assumptions relating to multiple correlation, see Ezekiel, Mordecai. "The Assumptions Implied in the Multiple Regression Equation," *Journal of American Statistical Association*, 20: 405-08, September, 1925.

In view of the various factors that may affect the dependability of an index of predictive efficiency, it is apparent that the calculated value should not be considered highly dependable unless one is able to show that the possible influences do not apply. The standard error of estimate and the derived indices of predictive efficiency are based upon a number of assumptions that are likely to be only roughly approximated when the regression equation is actually applied as a formula of prediction. In many cases, perhaps most cases, the indicated predictive efficiency will be greater than is actually realized. Hence, it is probably wise to discount somewhat the calculated index of predictive efficiency. A valid measure of predictive efficiency for a particular population may be obtained by deriving the regression equation and applying it to this population. The differences between the criterion measures and the predictions will be valid errors of estimate provided the criterion measures are accurate.

Practical minimum efficiency of prediction. An important question relates to the minimum efficiency for which prediction is justified as a practical procedure. The answer to this question will be a judgment, but in making the decision, one should clearly understand the measure of predictive efficiency employed, and bear in mind the use that is to be made of the predictions. In an article published in 1927 Hull ¹ expressed the opinion that when the efficiency (E_M) is less than 13 per cent, the value of making predictions is doubtful and that for efficiencies between 13 per cent and 20 per cent, the predictions are only "possibly useful." Hull's statement has been widely accepted, but it should be noted that he gives no indication of understanding the measure of efficiency of prediction (E_M) that he was dealing with. One can only speculate in regard to what his judgment might have been if he had understood that his measure of efficiency represented the per cent of improvement over using the mean as the estimate for each member of the population, and

¹ Hull, C. L. "The Coefficient of Correlation and Its Prognostic Significance," *Journal of Educational Research*, 15: 337, May, 1927.

that the per cent of improvement over chance prediction corresponding to E_M equal to .13 is 43. Furthermore, he did not consider the efficiency of predicting true measures of the criterion. Examination of Table XIV shows that when r_{00} is .80 or less, as is usually the case, the values of $E_{\infty g}$ are materially larger than those for E_g . These facts together with the interpretation of r_{01} in terms of per cent of predictions in error by more than a specified amount and especially the application of this interpretation procedure to the lower prognostic measures suggest that he might have arrived at a somewhat different opinion.

Although the calculated values of r_{01} and $R_{0.12\dots n}$ tend to exaggerate the predictive value of prognostic measures, the present writers are of the opinion that predictions may be sufficiently useful to justify the expense involved in calculating them when the correlation is as low as .40. In some situations predictions based upon even lower correlations may be worth while. They would, however, emphasize that predictions should always be used wisely. They do not subscribe to a pigeon-hole policy of placement in either educational or vocational guidance.

An explanation of the inefficiency of prediction.¹ In order to obtain a basis for considering means for improving prediction, it will be helpful to inquire into the causes of inefficiency. What we attempt to predict is commonly thought of as the resultant of a number of contributing factors or causes, and the use of the regression equation implies the assumption that the dependent variable x_0 is a linear function of these causes. If the causes are represented by ² $a_1, a_2, a_3 \dots a_m$, this assumption is expressed by writing

$$x_0 = c_{01}a_1 + c_{02}a_2 + c_{03}a_3 + \dots c_{0m}a_m$$

Although this assumption is not necessarily valid, it appears reasonable and may be accepted as an approximation. The cor-

¹ This explanation is in terms of an approach developed in the next chapter. If the reader experiences difficulty in understanding the explanation, he should return to it after studying Chapter XI.

² This argument is given in terms of deviation measures. This, however, does not restrict its application.

relation between the dependent variable and the independent variables and the intercorrelations of the latter may be explained by assuming a similar structure for the independent variables.¹ Hence, the variables of a multiple regression equation might have a structure similar to the following.

$$\begin{aligned}
 x_0 &= c_{01}a_1 + c_{02}a_2 + c_{03}a_3 + \dots + c_{0i}a_i + \dots + c_{0m}a_m \\
 x_1 &= c_{11}a_1 + c_{12}a_2 \qquad \qquad \qquad + c_{1s_1}a_{s_1} \\
 x_2 &= \qquad \qquad c_{22}a_2 + c_{23}a_3 + c_{2s_2}a_{s_2} \\
 x_3 &= c_{31}a_1 \qquad \qquad + c_{33}a_3 + c_{3s_3}a_{s_3} \\
 &\vdots \\
 x_n &= \qquad \qquad \qquad c_{n3}a_3 + c_{ni}a_i + c_{ns_n}a_{s_n}
 \end{aligned}$$

The causes $a_1, a_2, a_3 \dots a_i$ appear in one or more of the independent variables, but the remaining causes of the dependent variable are not found in any independent variable. Each independent variable includes a factor a_s that does not appear in the dependent variable. This factor may be merely the variable error of measurement, but usually it includes also a variable error of validity. It may include other elements not appearing in the dependent variable.

This analytical description of the variables involved in a regression equation offers a basis for pointing out the causes of inefficiency in prediction. In a typical situation several of the a 's (causes) of the dependent variable (x_0) are not found in any of the independent variables. For example, a student's success in first year of college is probably conditioned by his roommate and other companions, the kind and quality of instruction he receives, his instructors' methods and policies of grading, and other factors that are characteristics of the school community. It is also probably influenced by personal traits that are seldom measured. Since we are limited to those prognostic measures that may be obtained at the time the predictions are desired, it is obvious that estimates of success in college at the time of graduation from high school will inevitably involve errors of estimate due to the absence of a number of causes in the regres-

¹ See the explanation of factor analysis, pages 399 f.

sion equation. Similar statements may be made relative to other prediction situations.

A second cause of the inefficiency of predictions made from the regression equation is not so generally understood. A dependent variable cannot be precisely expressed as a linear function of a group of independent variables when they include factors such as indicated by the a 's with "s" subscripts not found in the dependent variable. The independent variables with which we deal in educational prediction usually include such factors and this condition contributes to the inefficiency of our prediction formula. In other words, if each cause of the dependent variable appeared in at least one of the independent variables, the regression equation would be only a best estimate of the relationship between the dependent variable and its causes.

A third cause of inefficiency in prediction is due to the overlapping of the independent variables. If we think of the dependent variable being expressed as a linear function of its elemental causes, a prognostic measure (independent variable) is likely to include two or more of these elemental causes, and a given elemental cause is likely to appear in two or more of the independent variables. The typical situation is probably similar to that represented by the following equations.

$$\begin{aligned}x_0 &= 3a_1 + 5a_2 + a_3 + a_m \\x_1 &= 2a_1 + 3a_2 + a_{s_1} \\x_2 &= 5a_2 + a_3 + a_{s_2} \\x_3 &= 3a_1 + a_2 + 6a_3 + a_{s_3}\end{aligned}$$

Due to the presence of a_{s_1} , a_{s_2} , and a_{s_3} it is very unlikely that the weighting of x_1 , x_2 , and x_3 in the regression equation will be such that $b_{01.23}(2a_1) + b_{03.12}(3a_1)$ will be equal to $3a_1$. Similar statements may be made with reference to the other elemental causes. Hence the overlapping of the independent variables contributes to the inefficiency of the predictions.

Securing the most accurate predictions for a given criterion measure. The three causes of inefficiency in prediction suggest two points of attack for decreasing the errors of estimate. In

the first place, one should attempt to secure prognostic measures that will include as many of the causes (a 's) of the dependent variable as possible. It appears, however, that this attempt cannot usually be completely successful because some of the causes cannot be measured at the time the predictions are desired. The other attack is to endeavor to secure "pure" measures of elemental causes. Factor analysis described in the following chapter appears to afford assistance in accomplishing this, but it would be hazardous to predict the degree of success that will eventually be attained. Random explorations to discover new prognostic measures are not likely to be very fruitful. The addition of an independent variable that includes no new "cause" will improve the prediction formula only very slightly if at all. The addition of an independent variable in which the causal elements are minor factors will also not result in much improvement.

Recognition of the causes of inefficiency affords an explanation of why in predicting success in the first year of college the increment of accuracy resulting from increasing the number of independent variables is not compatible with the additional labor involved.¹ In this connection it should be noted that the multiple correlation coefficient tends to exaggerate the accuracy of predictions.² In other words, if the multiple regression equation derived from one population is applied to a second population differing from the first only by chance, the estimated standard error of estimate will tend to be larger than the standard deviation of the actual errors of estimate. This phenomenon is referred to as the "shrinkage of the coefficient of multiple correlation." Two formulae³ have been proposed for calculating a value of R that will give a more accurate standard error of

¹ Hull, C. L. "The Joint Yield from Teams of Tests," *Journal of Educational Psychology*, 14: 396-406, October, 1923.

² See reference to Ezekiel on page 347.

³ One formula was derived by B. B. Smith and reported by M. J. B. Ezekiel at the meeting of the American Mathematical Society in 1928. The other formula was derived by Wherry. Both are given in Wherry, R. J. "A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation," *Annals of Mathematical Statistics*, 2: 440-57, 1931.

estimate. Larson ¹ has shown the shrinkage empirically. The increase of the coefficient of multiple correlation, due to adding one more variable, is small when the number of independent variables is already large. Hence, the shrinkage, due to the addition of a variable, may be greater than the increase. Larson found that for his data the shrinkage exceeded the increase when ten independent variables were included. In other words, slightly better predictions were obtained from eight variables than from ten variables.

ILLUSTRATIVE STUDIES OF PREDICTIONS

Since several studies of predicting success in first year in college and of predicting teaching success have been referred to in the preceding pages, the following references have been limited to other prediction situations.

GORDON, H. C. "The Specific Nature of Achievement and the Predictive Value of the IQ." A thesis submitted for the degree of Ph.D. in Education. Philadelphia: University of Pennsylvania, 1931. 147 pp.

The data for this study were obtained from the twelve senior high schools of Philadelphia. The value of the IQ as determined by the Otis Self-Administering Test of Mental Ability for predicting success in school subjects at the twelfth-grade level is studied critically. The report includes a summary and a carefully selected bibliography.

GROVER, C. C. "Results of an Experiment in Predicting Success in First Year Algebra in Two Oakland Junior High Schools," *Journal of Educational Psychology*, 23: 309-14, April, 1932.

The dependent variable is scores on the Columbia Research Bureau Algebra Test, and the two independent variables are scores on the Orleans Algebra Prognosis Test and intelligence quotients obtained by the Terman Group Test of Mental Ability. The correlation between the prognostic test and the measures of achievement in algebra is .61, while the multiple correlation coefficient is .65. The study may be recommended to the reader because of its clear description of the procedures employed.

KAULFERS, WALTER. "Value of English Marks in Predicting Foreign-Language Achievement," *School Review*, 37: 541-46, September, 1929.

In this study end-semester marks in English and average mid-semester and end-semester marks in high school Spanish were transmuted into point scores by means of the standard score procedure. Coefficients of correlation

¹ Larson, S. C. "The Shrinkage of the Coefficient of Multiple Correlation," *Journal of Educational Psychology*, 22: 45-55, January, 1931.

were then calculated of 54 boys and 55 girls. These coefficients are respectively .509 and .578.

KELLEY, T. L. "Educational Guidance, an Experimental Study in the Analysis and Prediction of Ability of High School Pupils," *Teachers College, Columbia University Contributions to Education*, No. 71. New York: Bureau of Publications, Teachers College, Columbia University, 1914. 116 pp.

In this pioneer study, Kelley investigated the predictive value of school marks in various subjects; teachers' estimates of intellectual ability, conscientiousness, emotional interest in school work, and oral expression; and special tests in algebra, English, history, geometry, and interests. Regression equations are reported for various combinations of these variables. In the appendix of the monograph are described techniques used in deriving comparable measures—one of the first applications of the standard score procedure.

LIMP, C. E. "The Use of the Regression Equation in Determining the Aptitude of an Individual," *Journal of Educational Psychology*, 16: 414-18, September, 1925. See also: Hull, C. L., and Limp, C. E. "The Differentiation of the Aptitudes of an Individual by Means of Test Batteries," *Journal of Educational Psychology*, 16: 73-88, February, 1925.

Difference between school marks in English and school marks in typewriting for the same individuals constitute the criterion measures, or dependent variable, in this study. The problem was that of predicting the difference between the two aptitudes rather than that of predicting aptitude in English or in typewriting.

MILLER, L. W. "An Experimental Study of the Iowa Placement Examinations," *University of Iowa Studies in Education*, Vol. 5, No. 6. Iowa City, Iowa: University of Iowa, 1930. 116 pp.

The prognostic efficiency of the Iowa Placement Examinations was studied in this research. The results presented include "inter-part correlations, reliability coefficients, correlation coefficients with first semester grades, means, standard deviations, and an item analysis for each part." Both ordinary multiple regression equations and multiple regression equations in the beta or standard score form are reported in the study. An excellent summary of previous research is given.

MORE, G. V. D. "Prognostic Testing in Music on the College Level: An Investigation Carried on at the North Carolina College for Women," *Journal of Educational Research*, 26: 199-212, November, 1932.

Intercorrelations for a group of 179 individuals were secured between average music marks and scores on a number of music tests, some of which

were devised by the author. A battery was selected from the tests having the highest correlations with the criterion. A coefficient of multiple correlation of .73 was obtained. The author presents an interesting interpretation of her findings.

PATERSON, D. G., et al. *Minnesota Mechanical Ability Tests*. Minneapolis: The University of Minnesota Press, 1930. 586 pp.

Seven of the tests tried out in the preliminary study yielded reliability and validity coefficients of sufficient magnitude to justify revision and further experimentation. The prognostic efficiency of batteries made up of various combinations of these tests was studied.

ROSS, C. C. "The Relation between Grade School Record and High School Achievement," *Teachers College, Columbia University Contributions to Education*, No. 166. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 70 pp.

In this comprehensive study, correlations are reported between average marks in various elementary school subjects and marks in various first-year high school subjects. Correlations are also reported between variables which represent composites of marks in several subjects and between high school achievement and other factors such as elementary school deportment, effort, and attendance. On page 42 is given an interesting graphic representation of the accuracy of prediction of general high school average, and achievement in Latin, English, and mathematics.

SEGEL, DAVID, and BRINTLE, S. L. "The Relation of Occupational Interest Scores as Measured by the Strong Interest Blank to Achievement Test Results and College Marks in Certain College Subject Groups," *Journal of Educational Research*, 27: 442-45, February, 1934.

Correlations between Strong Interest scores relative to engineering, medicine, law, life insurance, personnel management, and purchasing agent and measures of scholastic achievement in English, mathematics, science, history and social science, and measures of intelligence are reported in this study. While most of the coefficients are low, those for achievement in mathematics and science are high enough to be significant in the case of interest in engineering and in medicine. The authors advocate the use of interest schedules in educational guidance, but suggest the construction of one especially designed for this purpose.

TOORS, H. A. "Predicting the Returns from Questionnaires: A Study in the Utilization of Qualitative Data," *Journal of Experimental Education*, 3: 204-15, March, 1935.

An illustration of the possibilities of systematic prediction.

WEST, C. H. "The Practical Statistics of Prediction," *Journal of Experimental Education*, 3: 198-203, March, 1935.

Shows the effect of using a modification of the regression equation.

ZYVE, D. L. "A Test of Scientific Aptitude," *Journal of Educational Psychology*, 18: 525-46, November, 1927.

A test of scientific aptitude devised by the author was given to a group of 50 research students in physics, chemistry, and electrical engineering and the scores obtained were correlated with judgments of competent individuals respecting the same trait. Further study, and use of students in non-scientific departments indicated that the test "is a test of aptitude rather than of training, and is capable of differentiating scientific aptitude among highly selected and trained groups." In addition to reporting an ordinary regression equation, the author gives a regression equation for predicting "true" scores.

CHAPTER XI

IDENTIFYING AND STUDYING CAUSE AND EFFECT RELATIONSHIPS

The problems of this chapter. As indicated by the above title, this chapter deals with two problems, first, the identification of cause and effect relationships, and second, the measurement of the contributions from the causes in such relationships. In connection with the second problem, the techniques of correlation analysis will be described.

A. THE NATURE OF RELATIONSHIP

An explanation of relationship between variables. Two correlated variables may be described as being related. One may be a cause ¹ of the other or both may be effects of a common cause, or causes.² When the first condition prevails, the relationship is described as one of *cause and effect*. When the connection between the two variables is due to a common cause, the relationship is one of *concomitant variation*.

If x_0 and x_1 are two correlated variables expressed in deviation form, they may be thought of as being analyzable as follows:³

¹ Contemporary philosophical discussions reveal considerable controversy relative to the concept of causation. It will serve our purpose, however, to define a cause as an element in an existing situation which produces a condition different from that which otherwise would have prevailed. For an elaboration of this idea, see Lamprecht, S. P. "Causality," *Essays in Honor of John Dewey*. New York: Henry Holt and Company, 1929, p. 203.

² The principle that when two variables are correlated, one is the cause of the other or they are connected by a common cause was expressed by Mill several years before the idea of correlation as represented by a coefficient was developed by Galton.

Mill, J. S. *A System of Logic*. New York: Longmans, Green and Company, 1906, p. 263. This volume was first published in 1843.

³ For proof, see Kelley, T. L. *Crossroads in the Mind of Man*. Stanford University: Stanford University Press, 1928, p. 38. The statement of Kelley's second proposition is not identical with that given here but the proof is applicable.

$$\begin{aligned}x_0 &= c_0 a_{01} + b_0 \\x_1 &= c_1 a_{01} + b_1\end{aligned}$$

In these equations, c_0 and c_1 are constants, a_{01} is a variable but for any pair of values of x_0 and x_1 it has the same value and b_0 and b_1 are variables which are uncorrelated with each other and with a_{01} . In other words, this theorem states that if two variables are correlated, each one may be thought of as the weighted sum of two uncorrelated sub-variables, or factors, one of which is perfectly correlated with the corresponding factor of the other. The other sub-variables are uncorrelated with each other and with the *common factor*. Hence, the statement that two variables are related may be interpreted as meaning that they include a common factor, represented in the above analysis by a_{01} . This factor may not have the same name in both variables. Hence, the term "common" is not to be interpreted as meaning "the same," but rather as meaning that the values of a_{01} in one variable are numerically equivalent to the corresponding values in the other.

Two given correlated variables cannot be analyzed in the manner indicated ¹ but the analytical structure may be illustrated by using sums of uncorrelated variables. Table XV gives several illustrative values.²

Causal variables. The principle expressed on page 366 may be restated as follows: The common factor in one of two correlated variables may be a cause contributing to the other variable or the common factor in both variables may represent contributions

¹ The best estimate of $c_0 a_{01}$ is the regression of x_0 on x_1 , $r_{01} \frac{\sigma_0}{\sigma_1} x_1$. The best estimate of b_0 is the error of estimate $x_0 - r_{01} \frac{\sigma_0}{\sigma_1} x_1$ or $x_0 - \bar{x}_0$ which may be represented by $x_{0.1}$. Hence, we may write as the best estimate of the above analysis

$$\begin{aligned}x_0 &= \bar{x}_0 + x_{0.1} \\x_1 &= \bar{x}_1 + x_{1.0}\end{aligned}$$

² The values of the uncorrelated variables A_1 , A_2 , and A_3 were obtained by counting tosses of coins. For example, the values of A_1 were obtained by tossing thirty coins and counting heads or tails. In order to neutralize the effect of possible imperfections in the coins, the tails were counted for one-half the tosses.

from a common cause. Since measurement may be indirect, either variable may be thought of as representing indirect measures of the common factor or of the common cause. From this point of view, any relationship may be thought of as one of cause and effect, but this designation is reserved for those cases in which the name attached to one variable defines a cause contributing to the other. In other words, a causal variable is one whose name designates a cause of the specified effect. For example, the correlation between reading test scores and arithmetic test scores is not accepted as evidence that reading ability is a cause of arithmetical ability or vice versa. However, if one of the tests is designated as an instrument yielding measures of general intelligence, which would not be wholly unjustifiable, the relationship might be designated as one of cause and effect.

TABLE XV. ILLUSTRATIVE VALUES OF TWO CORRELATED VARIABLES
SHOWING THE COMMON FACTOR

$X_0 = A_1 + A_2$	$X_1 = A_1 + A_3$
35 = 15 + 20	24 = 15 + 9
32 = 18 + 14	29 = 18 + 11
27 = 15 + 12	28 = 15 + 13
26 = 16 + 10	27 = 16 + 11
33 = 18 + 15	25 = 18 + 7
33 = 18 + 15	32 = 18 + 14
29 = 19 + 10	33 = 19 + 14
30 = 17 + 13	31 = 17 + 14
28 = 15 + 13	26 = 15 + 11
28 = 11 + 17	25 = 11 + 14
23 = 12 + 11	25 = 12 + 13
21 = 14 + 7	26 = 14 + 12
33 = 18 + 15	32 = 18 + 14
35 = 17 + 18	27 = 17 + 10
25 = 13 + 12	22 = 13 + 9
33 = 18 + 15	26 = 18 + 8
31 = 18 + 13	39 = 18 + 21
33 = 16 + 17	24 = 16 + 8
23 = 13 + 10	21 = 13 + 8
26 = 13 + 13	25 = 13 + 12
29 = 17 + 12	31 = 17 + 14
28 = 14 + 14	23 = 14 + 9

The designation of a variable as a cause does not mean that it is wholly a cause. Usually it includes a factor (component) that is uncorrelated with the effect and reference to a variable as a cause should be interpreted to mean merely that it includes a factor that is a cause. Similarly, the designation of the variable as an effect should be thought of as meaning that it includes a factor that is the effect of a factor in the causal variable.

Techniques for identifying causes. Sometimes the problem of identifying a cause appears as one of determining the explanation of or the reasons for a certain condition or phenomenon. For example, an investigator may seek the explanation of failure or maladjustment in school. Such problems may be expressed in terms of the identification of causal variables, but there is little advantage in doing so.

The citation from Rice's study on page 271 illustrates the use of the comparative survey when the causal influence of a variable is being investigated. Another illustration is furnished by the study of Smillie and Spencer who measured the intelligence of five groups of pupils differing with respect to intensity of infestation with hookworm.¹ The data collected revealed that the heavier the hookworm infestation the lower the intelligence quotient, and they concluded that hookworm infestation is a cause of mental retardation. This procedure has been called the "method of differences."² The principle involved is that if two or more populations differ in respect to a given trait or characteristic the causes of this difference are to be found in the other traits or characteristic in which they also differ. The weakness of this method is that the identification is not certain. For example, in the investigation of Smillie and Spencer it may be that children of low intelligence are more likely to have hookworms than children of higher intelligence. Hence, it may be that the infestation is an effect rather than a cause of low intelligence or it may be that both are effects of a common cause.

¹ Smillie, W. G., and Spencer, C. R. "Mental Retardation in School Children Infested with Hookworms," *Journal of Educational Psychology*, 17: 314-21, May, 1926.

² Mill, *op. cit.*, pp. 255-60.

In applying this procedure as a means of identifying causes, an investigator should consider the other differences as merely possible causes and critically examine each before accepting it as the cause of a given trait or characteristic. If an investigator is critical and persistent, he will frequently be able to make a highly dependable identification. An excellent illustration of the reasoning involved is furnished by Brownell,¹ who sought an explanation of the difference in the performances of certain groups of children on an arithmetic test. A detail of the administration of the test, which at first was overlooked, turned out to be a clue to the explanation and hence the cause. It is relatively easy to collect a mass of data regarding differences, but the analysis and interpretation may require much critical thinking. Barr's study² illustrates the difficulties of interpretation. Although he secured much detailed information concerning the teaching performances, his conclusions relative to the causes of good teaching and poor teaching are admittedly not satisfactory.

As pointed out in Chapter IX, controlled experimentation is superior to the comparative survey, but it may not be a feasible procedure. If Smillie and Spencer had started with several equivalent groups of children who were free from hookworm and then inoculated the members of some of the groups, the resulting differences in intelligence, after a period of years, would be more dependable evidence than that obtained. This, however, would not be a defensible procedure.³ In other cases it would be difficult to introduce the desired change and maintain the experimental conditions for a sufficient period. For example, it would be difficult to ascertain the causes of good and poor teaching by controlled experimentation.

The "method of agreement"⁴ involves the survey of a group,

¹ Brownell, W. A. "An Evaluation of an Arithmetic 'Crutch,'" *Journal of Experimental Education*, 2: 5, September, 1933.

² Barr, A. S. *Characteristic Differences in the Teaching Performance of Good and Poor Teachers of the Social Studies*. Bloomington, Illinois: Public School Publishing Company, 1929. 127 pp.

³ This procedure was employed by Walter Reed in identifying the cause of yellow fever. The use of human subjects in this case was justified by the importance of the anticipated findings.

⁴ Mill, *op. cit.*, pp. 254-55.

the members of which have a certain trait or characteristic in common for the purpose of determining other common traits or characteristics. For example, in seeking the causes of pupil failure a group of such pupils might be surveyed to determine what other characteristics may be common to them or most of them. If this survey is highly detailed, it represents a series of case studies. The identification of other common traits or characteristics does not necessarily constitute identification of causes. The common traits or characteristics may be due to the operation of common causes.

If the correlation between a given effect and another variable is not zero, this fact is evidence that this variable is a possible cause. The existence of correlation, however, is not proof of causation. The correlation may be due to a common cause. For example, Ezekiel ¹ cites the illustration of the correlation between the number of automobiles passing a given point in Washington, D. C., during each fifteen minute period from noon until midnight and the height of the water in the Potomac River during the same periods. The height of the water in the Potomac River is affected by the ocean tides which are in turn influenced by the moon. The position of the sun has a definite influence on the movements of people, and at any given time there is a definite although complex relationship between the position of the sun and that of the moon. Hence, on certain days a significant correlation may be obtained between the numbers of automobiles and the heights of the water, but one cannot infer the underlying causal connection from the coefficient obtained. Hence, when employing correlation analysis in studying a problem of causation, it is necessary to demonstrate what variables are causes and the direction of causation by other means.

B. MEASUREMENT OF THE CONTRIBUTIONS

A measure of the contribution from a causal variable. Examination of Table XV reveals that the ratio of the values of the

¹ Ezekiel, Mordecai. *Methods of Correlation Analysis*. New York: John Wiley and Sons, 1930, p. 349.

common factor to the corresponding values of X_0 , the dependent variable is not constant. This condition suggests an "average" of the ratio of the common factor to X_0 , but since raw data are frequently expressed from arbitrary zero points, a measure based upon them will not be satisfactory. If the values of a variable are expressed as deviations from its mean, the standard deviation is a measure of its "magnitude" for a given population. From the point of view of this concept of the "magnitude" of a variable, the contribution of an independent variable may be thought of as the effect it produces upon the variability of the dependent variable within a given population. Since the "magnitude" of the common factor may also be described in terms of its variability, the problem of measuring the contribution of a causal variable may be thought of as involving the determination of a relation between the variability of the common factor and that of the dependent variable.

The standard deviation is a measure of the variability of a group of measures, but, in order to simplify some of the relations involved, σ^2 , called the *variance*, may be used as the measure of the "magnitude" of a variable. In terms of the symbolism on

page 367, the fraction $\frac{c_0^2 \sigma_{a_{01}}^2}{\sigma_0^2}$, called the *variance ratio*, gives the

per cent of the variance of the dependent variable that is due to the common factor. Hence, the problem of measuring the contribution from a causal variable within a given population is defined as that of determining the value of the variance ratio

$$\frac{c_0^2 \sigma_{a_{01}}^2}{\sigma_0^2}.$$

The value of the variance ratio.¹ The argument is simplified by expressing the two variables in standard score form.

$$z_0 = w_0 a_{01} + w_1 b_0$$

$$z_1 = v_0 a_{01} + v_1 b_1$$

¹ For a complete account of the argument of this theorem see Monroe, W. S. "Note on the Interpretation of Coefficients of Correlation" to be published in the *Journal of Educational Psychology*.

Since the standard deviation is the unit, $\sigma_0 = \sigma_1 = \sigma_{a_0} = \sigma_{b_0} = \sigma_{b_1} = 1.00$. If $w_0a_{01} + w_1b_0$ and $v_0a_{01} + v_1b_1$ are substituted for x_0 and x_1 in the formula, $r_{01} = \frac{\sum x_0x_1}{N\sigma_0\sigma_1}$, we obtain¹ $r_{01} = w_0v_0$.

The variance of w_0a_{01} is w_0^2 . Since the variance of z_0 is unity, w_0^2 is also the variance ratio. Similarly v_0^2 is the variance ratio which expresses the proportion of the variance of z_1 that is due to the common factor. By squaring the equation $r_{01} = w_0v_0$, we have $r_{01}^2 = w_0^2v_0^2$. Hence, we may state as a theorem: The square of the product-moment coefficient of correlation is equal to the product of the variance ratios of the two variables.

Since the variables are in terms of standard units, $w_0^2 + w_1^2 = 1.00$ and $v_0^2 + v_1^2 = 1.00$ and hence the limiting values of both w_0^2 and v_0^2 are 0.00 and 1.00. For a given coefficient of correlation, i.e., for a fixed value of $w_0^2v_0^2$, the minimum value of w_0^2 , the variance ratio of z_0 , occurs when $v_0^2 = 1.00$. The minimum value is r_{01}^2 . If $v_0^2 = w_0^2$, then $r_{01}^2 = (w_0^2)^2$ or $w_0^2 = r_{01}$. If $v_0^2 < w_0^2$, then w_0^2 is greater than r_{01} but the limiting value of w_0^2 is 1.00. Hence, the limits of the value of w_0^2 , the variance ratio of z_0 are r_{01}^2 and 1.00.

Significance of this theorem. This is a theorem of considerable importance when a coefficient of correlation is being interpreted in terms of the degree of communality of the two variables which is expressed by the variance ratio. The limits of the value of the variance ratio are r_{01}^2 and 1.00 and the value in a particular case is determined by the analytical structure of the two variables. For example, employing the symbolism of the preceding paragraphs, if v_1 is zero, $v_0^2 = 1.00$, and hence $w_0^2 = r_{01}^2$. This means that if the independent variable, z_1 , contributes itself completely, the value of the variance ratio is given by r_{01}^2 . If, on the other hand, the common factor in the independent variable causes only a small portion of its variability, i.e., if v_1 is relatively large in comparison with v_0 , then for the same value of

¹ The variables z_0 and z_1 may be analyzed into any number of factors. The coefficient of correlation will be the sum of the products of the coefficients of the common factors. This theorem is useful in theoretical work. See page 401.

r_{01} , w_0^2 , the variance ratio will have a value only slightly less than 1.00. In other words, for a given coefficient of correlation, say .50, the value of the variance ratio may be as small as .25 or as large as 1.00. The determination of its value within these limits depends upon the analytical structure of the two variables.

There is no general technique for determining the analytical structure of two correlated variables,¹ but in certain cases, an assumption appears reasonable. For example, in interpreting the correlation between general intelligence test scores and the scores on a general achievement test, Kelley² assumed that "that part of achievement which is not intelligence is as great an amount as that part of intelligence which is not achievement." Granting this assumption, which is not materially inconsistent with our understanding of the measures of these two traits, we have $w_1 = v_1$ and $r_{01} = w_0^2$. If in a given population the correlation between intelligence test scores and the scores on a general achievement test is .81, the value of the variance ratio is .81.

When chronological age is the independent variable, it appears reasonable to assume, at least in some cases, that it is contributed completely to the dependent variable. This means that in such cases the value of the variance ratio is r_{01}^2 . For example, if the correlation between the scores on an achievement test and chronological age is .40, this variable contributes only .16 of the variance of the achievement scores. On the other hand, in the case of measures of teaching success and intelligence test scores, for which the correlation may be taken as .16, it may be that the uncorrelated factor of the intelligence test scores is large in comparison with the common factor. If this is the case, the value of the variance ratio would be relatively large, possibly as large as .60. This would mean that only a

¹ Tryon has developed a method for calculating the value of the variance ratio, but since it requires two additional variables x_2 and x_3 defined so that a_{01} is the common factor for each pair of the four variables, the method is not generally applicable. Tryon, R. C. "The Interpretation of the Coefficient of Correlation," *Psychological Review*, 36: 423-24, September, 1929.

² Kelley, T. L. *Interpretation of Educational Measurements*. Yonkers-on-Hudson: World Book Company, 1927, p. 195.

minor factor of general intelligence, as measured, contributes to teaching success but this factor produces three-fifths of the variance of the measures of teaching success.¹

Effect of a factor of heterogeneity upon the relation between two variables. If a third variable is added to two paired variables,² the correlation between them will be affected. For example, suppose

$$\begin{aligned}x_0 &= a_2 + a_3 \\x_1 &= a_2 + a_4\end{aligned}$$

If a_1 is added to both variables, the population is described as heterogeneous with reference to $x_2 = a_1$ and $r_{(x_0 + a_1)(x_1 + a_1)}$ will be different from r_{01} . In other words, the coefficient of correlation between two variables depends upon the heterogeneity of the population with reference to other correlated variables. The effect of a factor of heterogeneity may be illustrated by considering the correlation between achievement test scores as measures of the dependent variable and intelligence test scores as measures of the independent variable. For a population of pupils belonging to the same school grade, the value of r_{01} is frequently less than .50,³ but if the population is taken from a sequence of grade groups, the value of r_{01} for scores yielded by the same tests will be materially greater. For a sequence of six grades, it may be as much as .80 or .90.

The effect of a third variable (factor of heterogeneity) that is positively or negatively correlated with both x_0 and x_1 is to

¹ Since in the standard score equations $w^2 + w_1^2 = 1.00$ and $v_0^2 + v_1^2 = 1.00$, factor patterns may be written for a given coefficient of correlation. For example, in the above illustration we might write

$$\begin{aligned}z_0 &= \sqrt{.64}a_{01} + \sqrt{.36}b_0 \\z_1 &= \sqrt{.04}a_{01} + \sqrt{.96}b_1 \\r_{01} &= \sqrt{.64}\sqrt{.04} = .16 \quad w_0^2 = .64\end{aligned}$$

The reader will find it illuminating to write the factor patterns corresponding to different assumptions in regard to the structure of two correlated variables.

² The simplified symbolism used here and in several of the following pages is introduced as a matter of convenience. Nothing would be gained by expressing a variable as a weighted sum of its components. The component variables a_1, a_2, a_3 , etc., are uncorrelated with each other.

³ The value of r_{01} depends upon the tests administered.

increase the "magnitude" of the common component a_{01} . If the factor of heterogeneity is positively correlated with one of the variables and negatively with the other, its effect is to decrease the correlation between the two variables. This case, however, is not often encountered in educational research and usually the effect of heterogeneity is to cause the obtained coefficient to be larger than the one for the corresponding homogeneous population.

A supplementary statement in regard to the meaning of the contribution of a causal variable. The preceding explanation of the effect of the factor of heterogeneity upon the common factor of a relationship emphasizes that a question concerning the contribution of a causal variable is indefinite until a population is specified. The variance ratio for a heterogeneous population measures a composite of the contribution of the causal variable and the contribution from the factor or factors of heterogeneity. Hence, it tends to be misleading to speak of the contribution of a given causal variable in a population that is heterogeneous with reference to one or more factors. The terminology may be used as a means of convenience, but an investigator should bear in mind the nature of the contribution when interpreting his findings.

The fact that the obtained measure of a contribution is for a particular population suggests the desirability of agreeing upon one or more standard populations. McCall, Kelley, and others have proposed that in dealing with certain problems of test construction an unselected group of twelve-year-old children be used as the standard population.¹ The adoption of this population as standard would be helpful in dealing with some problems, but frequently we desire a measure of the "net" contribution of a causal variable, that is, a measure of its contribution in a population that is homogeneous with respect

¹ Kelley has presented data in support of the thesis that a population made up of equal unselected groups from six consecutive grades is approximately equal to that of an unselected age group.

Kelley, T. L. *Interpretation of Educational Measurements*. Yonkers-on-Hudson: World Book Company, 1927, pp. 197 f.

to one or more other variables. Since collecting data from a homogeneous population would frequently involve obvious difficulties, it is appropriate to consider how the coefficient of correlation for a desired population may be estimated from the data collected from a heterogeneous population.

C. PARTIAL CORRELATION

Estimating the coefficient of correlation for a population homogeneous with respect to related variables—partial correlation.¹ The operation of the partial correlation technique may be illustrated by referring to a study by Terman.² The correlation between depth of chest and mental age for a large group of gifted boys ranging in age from 9 to 14 years is given as +.582. This population is heterogeneous with reference to chronological age. Its correlations³ with depth of chest and with mental age are given as +.618 and +.941. The partial correlation formula for three variables is

$$r_{01.2} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1 - r_{02}^2}\sqrt{1 - r_{12}^2}}$$

Let the symbols X_0 represent depth of chest, X_1 represent mental age, and X_2 represent chronological age. Substituting the given values

$$r_{01.2} = +.002$$

This value indicates that in a population⁴ homogeneous with respect to chronological age there is no relationship between depth of chest and mental age. This conclusion is in agreement with a priori reasoning.

¹ May has derived a more general technique, which is applicable even when the factor of heterogeneity is sex, nationality, or some other unordered series.

May, M. A. "A Method for Correcting Coefficients of Correlations for Heterogeneity in the Data," *Journal of Educational Psychology*, 20: 417-23, September, 1929.

² Terman, L. M., et al. *Genetic Studies of Genius*, Vol. I. Stanford, California: Stanford University Press, 1925, p. 168.

³ *Ibid.*, p. 156 and p. 168.

⁴ The age level of this homogeneous population is not specified. The value +.002 may be thought of as an "average" of the coefficients of correlation for the age levels included in the population from which the data were obtained.

In the same source, the correlation between standing height and mental age for the same group of gifted boys is reported as .835, between standing height and chronological age as .845, and between mental age and chronological age as .941. Noting that with respect to the partial correlation formula $r_{01} = +.835$, $r_{02} = +.845$, and $r_{12} = +.941$, we have

$$r_{01.2} = +.220$$

The value $+.220$ indicates a slight relationship between standing height and mental age for the group of gifted boys when the effect of the common cause, chronological age, has been eliminated. One should not infer, however, that mental age is a cause of standing height or that standing height is a cause of mental age. One important common cause has been eliminated, but it is possible that much of the correlation represented by the partial correlation coefficient $+.220$ is due to other common causes, for example, deep seated physiological factors which may affect both height and intellect.

When the effects of two variables, x_2 and x_3 , are to be partialled out, there are two formulae. Either may be employed but by substituting the indicated values in both, a check upon the calculations is obtained.

$$r_{01.23} = \frac{r_{01.2} - r_{03.2}r_{13.2}}{\sqrt{1 - r_{03.2}^2} \sqrt{1 - r_{13.2}^2}}$$

$$r_{01.32} = \frac{r_{01.3} - r_{02.3}r_{12.3}}{\sqrt{1 - r_{02.3}^2} \sqrt{1 - r_{12.3}^2}}$$

Ordinary coefficients of correlation are called *zero-order* coefficients. When one variable has been eliminated, as for example, $r_{01.3}$, the coefficient is designated as one of the *first-order*. It will be observed that first-order coefficients are substituted in the above equations to obtain one of the *second-order*. When three or more variables are eliminated, the formulae may be written in an increasing number of ways to secure

checks. The general formula for partial correlation for the elimination of $n - 1$ independent variables ¹ is as follows:

$$r_{01.23\dots n} = \frac{r_{01.23\dots(n-1)} - r_{0n.23\dots(n-1)}r_{1n.23\dots(n-1)}}{\sqrt{1 - r_{0n.23\dots(n-1)}^2}\sqrt{1 - r_{1n.23\dots(n-1)}^2}}$$

When the number of variables to be partialled out is greater than one, the calculation is laborious, but as in the case of multiple regression described in the preceding chapter, a number of time saving techniques and aids have been devised. The tables by Holzinger ² and Miner ³ provide the square roots required in the calculations. Holzinger ⁴ describes a method of calculating partial correlation coefficients in which determinants are used. The devices by Hull ⁵ and Wood ⁶ are also useful. Baten has prepared tables for finding partial coefficients.⁷

A precise definition of partial correlation. The partial correlation technique is commonly described as a means of eliminating the effect of one or more other variables (factors of heterogeneity) from a relationship or as a means of securing the coefficient of correlation in a population for which the one or more other variables are constant. A more accurate description is that the coefficient of partial correlation is the correlation between the residuals formed by subtracting the regressions of x_0 and x_1 on the variable or variables partialled out.⁸ For exam-

¹ The reader should note that according to the symbolism used n is the number of independent variables. The total number of variables is $n + 1$.

² Holzinger, K. J. *Statistical Tables for Students in Education and Psychology*. Chicago: University of Chicago Press, 1925. 74 pp.

³ Miner, J. R. *Tables of $\sqrt{1 - r^2}$ and $1 - r^2$ for Use in Partial Correlation and in Trigonometry*. Baltimore: Johns Hopkins University Press, 1922. 50 pp.

⁴ Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, pp. 312-15.

⁵ Hull, C. L. "A Device for Determining Coefficients of Partial Correlation," *Psychological Review*, 28: 377-83, September, 1921.

⁶ Wood, E. R. "A Graphic Method of Obtaining the Partial-Correlation Coefficients and the Partial Regression Coefficients of Three or More Variables," *Supplementary Educational Monographs*, No. 37. Chicago: University of Chicago Press, 1931. 72 pp. The charts described in this monograph are published by E. R. Wood, State Department of Education, Columbus, Ohio.

⁷ Baten, W. D. "Tables for Finding the Partial Coefficient of Correlation," *Journal of Experimental Education*, 3: 170-73, March, 1935.

⁸ For the proof of this statement, see Dunlap, J. W., and Cureton, E. E. "On

ple, $r_{01.2}$ is the coefficient of correlation between the residuals, $x_{0.2} = x_0 - r_{02} \frac{\sigma_0}{\sigma_2} x_2$ and $x_{1.2} = x_1 - r_{12} \frac{\sigma_1}{\sigma_2} x_2$. When two variables are partialled out, the residuals are $x_{0.23} = x_0 - b_{02.3}x_2 - b_{03.2}x_3$ and $x_{1.23} = x_1 - b_{12.3}x_2 - b_{13.2}x_3$. In general, $r_{01.23 \dots n}$ is the correlation between $x_{0.23 \dots n}$ and $x_{1.23 \dots n}$.

Dependability of coefficients of partial correlation. The definition of the coefficient of partial correlation as the correlation between residuals affords a basis for considering the dependability of this statistic when it is labeled the coefficient of net correlation; that is, the coefficient of correlation between two variables when the effect of one or more other variable is eliminated or when one or more other variables are held constant. The usual interpretation of a coefficient of partial correlation implies that the factor pattern is of the type

$$\begin{aligned}x_0 &= a_1 + a_2 + a_4 \\x_1 &= a_1 + a_2 + a_3 \\x_2 &= a_1\end{aligned}$$

The residuals are $x_{0.2} = (a_1 + a_2 + a_4) - r_{02} \frac{\sigma_0}{\sigma_2} a_1$

$$x_{1.2} = (a_1 + a_2 + a_3) - r_{12} \frac{\sigma_1}{\sigma_2} a_1$$

In this case, which may be described as the one in which the variable partialled out is a component of the other two, $r_{02} \frac{\sigma_0}{\sigma_2}$ and $r_{12} \frac{\sigma_1}{\sigma_2}$ are equal to unity¹ and the residuals are formed by

the Analysis of Causation," *Journal of Educational Psychology*, 21: 664-65, December, 1930.

The concept of partial correlation as the correlation between residuals (errors of estimate) appears in Yule's treatment of the topic, but subsequent writers do not appear to have carried the idea over to the interpretation of the results of partial correlation. Yule, G. U. *An Introduction to the Theory of Statistics*, Eighth Edition. London: Charles Griffin and Company, Ltd., 1927, pp. 233 f.

¹ This follows from the fact that when x_2 is a component of x_0 , $r_{02}^2 = \frac{\sigma_2^2}{\sigma_0^2}$.

This condition implies equivalent units. In case the units are not equivalent, the values of the two expressions will need to be weighted to compensate for the inequality.

subtracting x_2 from both x_0 and x_1 . Hence, when the factor pattern is of this type, partial correlation does yield the desired net correlation.

If x_2 is not a component of x_0 and x_1 , $r_{02} \frac{\sigma_0}{\sigma_2}$ and $r_{12} \frac{\sigma_1}{\sigma_2}$ will not be unity and the effect of x_2 upon x_0 and x_1 will not be wholly eliminated. In addition, new factors of heterogeneity will be introduced. Hence, the value of $r_{01.2}$ will not be what is desired. It will be only a best estimate.

The operation of partial correlation can be illustrated by using variables constructed from uncorrelated components such as may be obtained by counting tosses of coins.¹ For example, suppose

$$\begin{aligned} X_0 &= A_1 + A_2 + A_3 + A_4 + A_5 \\ X_1 &= A_1 + A_2 + A_6 \end{aligned}$$

The coefficient of correlation $r_{01} = .630$. If $X_2 = A_1$, it is a component of X_0 and X_1 . Application of the partial correlation technique gives $r_{01.2} = .500$, which is approximately the value of the coefficient of correlation² between $X_{0.2} = A_2 + A_3 + A_4 + A_5$ and $X_{1.2} = A_2 + A_6$. If instead, $X_2 = A_1 + A_4 + A_7$, $r_{01.2} = .55$ which shows that the correlation obtained is not that for X_0 and X_1 in a population homogeneous with respect to X_2 . As another illustration, suppose

$$\begin{aligned} X_0 &= A_1 + A_2 + A_6 \\ X_1 &= A_1 + A_3 + A_7 \end{aligned}$$

If $X_2 = A_1$ is partialled out, $r_{01.2} = -.025$ which is the coefficient of correlation between $A_2 + A_6$ and $A_3 + A_7$. If instead,

¹ See page 367 for explanation of procedure. In the illustration here and on the following pages $N = 100$. The number of coins tossed for the component variables were as follows: A_1 , 30 coins; A_2 , 28 coins; A_3 , 24 coins; A_4 , 20 coins; A_5 , 16 coins; A_6 , 16 coins; A_7 , 16 coins. These calculations and others from similar data in the following pages are from Stuit, D. B. "Correlation Analysis as a Means of Solving Problems of Functional Relationship." A thesis submitted for the degree of Ph.D. in Education. Urbana: University of Illinois, 1934. 109 pp.

² The calculated value is .502. The slight discrepancy is due mainly to the fact that some of the intercorrelations between the component variables obtained from 100 tosses of coins are not precisely zero.

$X_2 = A_1 + A_4 + A_7$ is partialled out, $r_{01.2} = .23$. This is the coefficient of correlation between $[(A_1 + A_2 + A_6) - r_{02} \frac{\sigma_0}{\sigma_2}(A_1 + A_4 + A_7)]$ and $[(A_1 + A_3 + A_7) - r_{12} \frac{\sigma_1}{\sigma_2}(A_1 + A_4 + A_7)]$.

Since neither $r_{02} \frac{\sigma_0}{\sigma_2}$ nor $r_{12} \frac{\sigma_1}{\sigma_2}$ is equal to 1.00, both residuals contain a fractional multiple of A_1 as a component. Hence, the variables defined by these residuals are not perfectly homogeneous with respect to A_1 . In addition, the variable defined by the first difference has been made heterogeneous with respect to A_4 and A_7 and the variable defined by the second difference has been made heterogeneous with respect to A_4 . Hence, if the purpose was to partial out A_1 , it has not been accomplished in a satisfactory manner.

Types of factor patterns in educational research. There is no technique for determining the factor pattern of three inter-correlated variables, but our experience suggests certain types for certain cases. Chronological age appears to approximate a component of mental age and of a number of other variables.¹ Hence, when chronological age is to be partialled out, we may usually assume a factor pattern of the type illustrated on page 380. It appears that few, if any, other variables may be regarded as components. Test scores always involve variable errors of measurement as a factor and they are likely to include one or more other factors that do not appear in the other two variables. Hence, the factor pattern in educational research is likely to be one of the more complex types.

¹ Beyond certain limits the correlation may not be linear. Dunlap and Cureton suggest that where this condition prevails, chronological age may be eliminated satisfactorily by partialing out as separate variables chronological age, chronological age squared, chronological age cubed, and so on. They state that it is usually unnecessary to go beyond the cubes, and in a great many cases not beyond the squares.

Dunlap, J. W., and Cureton, E. E. "On the Analysis of Causation," *Journal of Educational Psychology*, 21: 663, December, 1930. Dunlap and Cureton refer in this connection to the work of Fisher on time series which are essentially similar to chronological age.

Fisher, R. A. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925, p. 172.

Semi-partial correlation. By means of a modification of the technique described in the preceding pages, the partialing out process may be applied to only one of the two correlated variables. By applying this modified technique, which is known as semi-partial correlation,¹ we may obtain the correlation between

$(x_0 - r_{02} \frac{\sigma_0}{\sigma_2} x_2)$ and x_1 . The formula for accomplishing this is

$$r_{(0.2)1} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1 - r_{02}^2}}$$

As an illustration of the application of semi-partial correlation, Dunlap and Cureton cite the problem of investigating the

¹ For derivation and discussion of semi-partial correlation for any number of variables, see

Dunlap, J. W., and Cureton, E. E. "On the Analysis of Causation," *Journal of Educational Psychology*, 21: 663-72, December, 1930.

From their general development, Dunlap and Cureton derive formulae for certain other systems of semi-partial correlation.

The original contribution of semi-partial correlation should be credited to Spearman, although his formula for three variables differs slightly from that of Dunlap and Cureton. The present writers are grateful to Dr. B. S. Burks for calling their attention to this point. See

Spearman, C. "The Proof and Measurement of Association between Two Things," *American Journal of Psychology*, 15: 94, 1904.

Franzen also proposed a formula for semi-partial correlation prior to the contribution of Dunlap and Cureton. It is identical, except in symbols, with their formula for three variables. See

Franzen, Raymond. "A Comment on Partial Correlation," *Journal of Educational Psychology*, 19: 194-97, March, 1928.

A reader who is making an intensive study of correlation analysis should become familiar also with part correlation which was devised by B. B. Smith in collaboration with Mordecai Ezekiel. The first account of this technique was published in "Correlation Theory and Method Applied to Agricultural Research," mimeographed publication, Bureau of Agricultural Economics, United States Department of Agriculture, August, 1926, pp. 57-60. Derivation and explanation are given in:

Ezekiel, Mordecai. *Methods of Correlation Analysis*. New York: John Wiley and Sons, 1930, pp. 181-84, 379-80.

The part correlation of x_0 with x_1 is the correlation between x_1 and the residual formed by subtracting the terms of the multiple regression equation, except that containing x_1 , from x_0 . Thus, for the case of two independent variables, the coefficient of part correlation is the correlation between $(x_0 - b_{02.1} x_2)$ and x_1 . In the case of semi-partial correlation, the residual would be $x_0 - b_{02} x_2$. The coefficient of part correlation, $r_{01.2}$, represents the correlation between variable x_1 and the residue of variable x_0 after eliminating an estimate of the contribution of the part of variable x_2 which is independent of x_1 . The coefficient of semi-partial correlation, $r_{(0.2)1}$, represents the correlation between variable x_1 and the residue of variable x_0 after eliminating an estimate of the contribution of all of variable x_2 .

possibility of home environment contributing to the intelligence test scores of children.¹ Obviously the intelligence of parents will contribute to home environment, the more intelligent parents providing the better environment. Hence, we shall expect a moderate degree of correlation between child intelligence test scores and measures of home environment due to the operation of parental intelligence as a common cause. Partial correlation is not satisfactory for dealing with this problem because measures of parental intelligence would be partialled out from both child intelligence test scores and measures of home environment. Assuming that parental intelligence as measured is a component of both the measures of environment and of the measures of child intelligence,² the net correlation obtained by means of partial correlation would be between that part of child intelligence which is independent of parental intelligence and that part of environment which is independent of parental intelligence. We are interested, however, in the correlation between all of child intelligence and that part of environment which is independent of parental intelligence. In other words, we desire the correlation between child intelligence and environment in an environment that is homogeneous with reference to parental intelligence. Semi-partial correlation provides a technique for partialing out measures of parental intelligence from only the measures of environment. This technique, however, will not be satisfactory unless the measures of parental intelligence are completely contributed to the measures of environment.

Interpretation of coefficients of correlation when the relationship is not one of cause and effect.³ Although the variance ratio and its relation to the coefficient of correlation have been

¹ The reader who consults the reference by Dunlap and Cureton should note that their system of symbolism is not identical with that used in the formula given here.

² This assumption probably is not true.

³ The interpretation of the coefficient of correlation as an index of probable errors of measurement was treated in Chapter VII and as an index of predictive efficiency in Chapter X. See also the discussion of the coefficient of correlation in Chapter IV.

introduced as techniques for the study of cause and effect relationships, they are applicable in other situations. A test score may be thought of as the algebraic sum of the true score and the variable errors of measurement. In terms of symbols

$$X_1 = X_\infty + e_1$$

A score on a duplicate form of the test will be represented by

$$X_I = X_\infty + e_I$$

It may be assumed that $\sigma_{e_1} = \sigma_{e_I}$. Hence,

$$r_{1I} = \frac{\sigma_\infty^2}{\sigma_1^2} = \frac{\sigma_\infty^2}{\sigma_\infty^2 + \sigma_e^2}$$

This relationship may be expressed by saying that the per cent of the variance of a group of test scores due to the trait measured by the test is given by the coefficient of reliability.¹ It may also be stated that the per cent of the variance due to the variable errors of measurement is given by $1 - r_{1I}$. For example, when the coefficient of reliability is .90, then 90 per cent of the variance of the test scores is due to the trait measured by the test and 10 per cent of it is due to the variable errors of measurement.

When a new test is constructed, the correlation of the scores yielded by it with the criterion measures is calculated as an index of its validity. In the typical case, the ratio of the variance of the common factor to the variance of the scores yielded by the test is not greatly less than the coefficient of validity. For example, suppose a coefficient of validity of .80 is obtained for a given test. Then approximately 80 per cent of the variance of the test scores is due to the common factor, i.e., to the factor of the criterion that the test measures. Less precisely, it may be said that on the average the test measures 80 per cent of what it is intended to measure.

This interpretation of a coefficient of validity should not be made unless it is clearly understood. In the first place, it is

¹ It should be noted that the coefficient of reliability must be for the population being considered.

applicable only to the population represented by the data for which the coefficient of validity was calculated. Furthermore, the statement that "on the average the test measures 80 per cent of what it is intended to measure" is appropriate only when the scores are expressed in deviation form. When the test is designated to measure achievement, the interpretation is subject to a further limitation. In most subject-matter fields, there is a fairly high correlation between the scores yielded by an achievement test and those yielded by an intelligence test. This means that a factor of what we measure as achievement is also measured by an intelligence test. A similar statement may be made with reference to criterion measures. Hence, it seems reasonable to say that a portion of the factor common to the achievement test scores and the criterion measures consists of general intelligence. If achievement is defined as the product of learning, that is, as "net achievement" which is implied when gains are computed,¹ the interpretation suggested in the preceding paragraph is not appropriate. In other words, the suggested interpretation of a coefficient of validity is not applicable in the case of gains in achievement or when the meaning of "net achievement" is intended.

This point may be illustrated noting the probable nature of the factor pattern of the test scores and of the criterion measures. Let a_1 represent the common factor contributed by intelligence, a_2 the remainder of the common factor, a_3 the specific factor of the test scores, a_4 the remainder of the net achievement as defined by the criterion, and e_0 and e_1 the respective variable errors of measurement.

$$\begin{aligned}x_0 &= a_1 + a_2 + a_3 + e_0 \\x_1 &= a_1 + a_2 + a_4 + e_1\end{aligned}$$

It is apparent that r_{01} would be greater than $r_{(x_0 - a_1)(x_1 - a_1)}$ which would be an index of the portion of the "net achievement" measured by the test. Hence, the value of r_{01} will exaggerate the validity of the test when it is thought of as measuring the prod-

¹ See page 303.

uct of learning. A more dependable index of the validity would be obtained if the contribution of general intelligence could be partialled out, but from our knowledge of the nature of intelligence tests, it is apparent that this cannot be done satisfactorily. Partialing out general intelligence may, however, indicate the extent to which the calculated coefficient of validity should be discounted. For example, Wood ¹ reported a correlation of .605 between the scores yielded by a true-false test and examination grades. The correlations with intelligence test scores were .451 and .386. The coefficient of partial correlation is .523. This value is probably somewhat in excess of the actual correlation between the true-false test scores and the examination grades after the effect of general intelligence is eliminated. But it indicates that the reported coefficient of validity of .605 should be rather heavily discounted when an interpretation is made relative to the community of function of the true-false test and the essay examination exclusive of general intelligence.

The hazards of applying partial correlation in such cases may be indicated by employing correlations reported by Corey.² He gives coefficients of correlation corrected for attenuation as follows:

New Type Test and Essay Examination	.93
New Type Test and Army Alpha	.62
Essay Examination and Army Alpha	.39

When intelligence is partialled out, the correlation between new type test scores and essay examination grades is .95 which does not seem reasonable:

The need for critical thinking in interpreting partial correlation coefficients is apparent. The promiscuous calculation of coefficients of partial correlation and the mechanical interpretation of them without considering the probable nature of the

¹ Wood, B. D. *Measurement in Higher Education*. Yonkers-on-Hudson: World Book Company, 1923, p. 188.

² Corey, S. M. "The Correlation between New-Type and Essay Examination Scores and the Relationship between Them and Intelligence as Measured by Army Alpha," *School and Society*, 32: 849-50, December, 1930.

factor patterns involved have mislead many investigators. It should be emphasized that partial correlation accomplishes what it is commonly considered to accomplish only when the factor pattern is of the type illustrated on page 380.

When dealing with data in the form of ratios such as IQ's, achievement quotients, or per cents, it is necessary to recognize what is known as "spurious correlation." If an intelligence test and an achievement test are administered to a group of children heterogeneous with reference to chronological age and the correlation between the obtained scores is zero, which would probably be the case if the reliabilities of the test were zero, the correlation between the IQ's and AQ's would be .50. This correlation is not without meaning, but care must be exercised in interpreting it or any coefficient of correlation obtained from ratios. The calculation of ratios introduces a common component or increases the one in the original measures.¹

Partial correlation as a means of identifying cause and effect relationships. On page 371, it was emphasized that the existence of correlation between two sets of paired measures is not proof that one variable is a cause of the other. The relationship may be due to the operation of a common cause. When it is effective, partial correlation removes the contribution of a common cause from the common factor. This suggests that if all factors of heterogeneity were partialled out, the resulting coefficient of partial correlation would be evidence of a cause and effect relationship between the two original variables. Theoretically this argument is defensible, but in practice a coefficient of partial correlation should not be interpreted as evidence of a cause and effect relationship because this statistic is not dependable un-

¹ Yule, G. U. *An Introduction to the Theory of Statistics*. London: Charles Griffin and Co., Ltd., 1917, pp. 214-15.

Holzinger, K. J. "Formulas for the Correlation between Ratios," *Journal of Educational Psychology*, 14: 344-46, September, 1923.

Thomson, G. H., and Pintner, Rudolph. "Spurious Correlation and Relationship between Tests," *Journal of Educational Psychology*, 15: 433-44, October, 1924.

For additional references on this topic, see Walker, Helen M. *Studies in the History of Statistical Method*. Baltimore: Williams and Wilkins Company, 1929, p. 124.

less the variable partialled out is a component of the other two, and because it is not possible to determine when *all* factors of heterogeneity have been included. It is exceedingly unfortunate that a number of writers have asserted or implied that partial correlation may be employed to identify cause and effect relationships.¹

D. REGRESSION EQUATIONS AND FACTOR ANALYSIS

Contributions of two or more independent variables. A dependent variable may be thought of as the sum of the contributions of its causes. A number of persons have interpreted the coefficients of the multiple regression equation as measuring these contributions. A case that has attracted considerable attention is that of the following equation reported by Burt.²

$$\text{Binet} = .54 \text{ School Work} + .33 \text{ Intelligence} + .11 \text{ Age}$$

In interpreting this equation Burt says that

of the gross result (mental age score on Binet test), then, one-ninth is attributable to age, one-third to intellectual development, and over one-half to school attainment. School attainment is thus the preponderant contributor to the Binet-Simon tests. To school the weight assigned is nearly double that of intelligence alone, and distinctly more than that of intelligence and age combined. *In determining the child's performance in the Binet-Simon Scale, intelligence can bestow but little more than half the share of the school, and age but one-third the share of intelligence.*

This statement has been criticized by Holzinger and Freeman.³ In the first place the regression equation does not prove the

¹ For example, see

McCall, W. A. *How to Experiment in Education*. New York: The Macmillan Company, 1923, p. 239.

Reavis, G. H. "Factors Controlling Attendance in Rural Schools," *Teachers College, Columbia University Contributions to Education*, No. 108. New York: Bureau of Publications, Teachers College, Columbia University, 1920. 69 pp.

² Burt, C. *Mental and Scholastic Tests*. London: P. S. King and Son, 1921, p. 183.

³ Holzinger, K. J., and Freeman, F. N. "The Interpretation of Burt's Regres-

existence of a cause and effect relationship between the variables of the equation. The regression equation may properly be thought of as representing a means of predicting mental age as measured by the Binet test, but this concept is very different from thinking of the variables considered independent as representing causes contributing to the dependent variable as an effect.¹ However, if the assumption of a cause and effect relationship is granted, Burt's interpretation of the regression coefficients is still not justified.

As an introduction to an appropriate interpretation, consider the case in which $x_0 = x_1 + x_2$, the independent variables being uncorrelated. Since the variables are expressed as deviations from their respective means,

$$\begin{aligned}\sigma_0^2 &= \frac{\Sigma x_0^2}{N} = \frac{\Sigma (x_1 + x_2)^2}{N} \\ &= \frac{\Sigma (x_1^2 + 2x_1x_2 + x_2^2)}{N} \\ &= \frac{\Sigma x_1^2}{N} + \frac{2\Sigma x_1x_2}{N} + \frac{\Sigma x_2^2}{N}\end{aligned}$$

Since x_1 and x_2 are uncorrelated, $\Sigma x_1x_2 = 0$ and we have $\sigma_0^2 = \sigma_1^2 + \sigma_2^2$. Hence, the per cent of the variance of x_0 con-

sion Equation," *Journal of Educational Psychology*, 16: 577-82, December, 1925.

The criticism by these writers is agreed with by Thomson. His article and the rejoinder by Holzinger and Freeman are of interest in this connection.

Thomson, G. H. "The Interpretation of Burt's Regression Equation," *Journal of Educational Psychology*, 17: 301-09, May, 1926.

Holzinger, K. J., and Freeman, F. N. "Rejoinder on Burt's Regression Equation," *Journal of Educational Psychology*, 17: 384-86, May, 1926.

¹ The fact that a regression equation is not proof of the existence of a cause and effect relationship should be emphasized. The use of the verb "contribute" in referring to the relation between the variables of a regression equation suggests, if it does not definitely imply, a cause and effect relationship. Sometimes the writer is probably aware that the assumption of such relationship is not justified and does not intend to imply it. Even when this is obviously the case, the uncritical reader is likely to be misled. For example, this is likely to happen in reading Garret's discussion on pages 256-57 of his text, *Statistics in Psychology and Education*. Unfortunately some writers appear to think of the regression equation as evidence of a cause and effect relationship. Some illustrations are to be found in the studies of the factors associated with teaching success. See pages 353-54 for references.

tributed by x_1 is given by the ratio $\frac{\sigma_1^2}{\sigma_0^2}$. Similarly, $\frac{\sigma_2^2}{\sigma_0^2}$ gives the per cent contributed by x_2 . The value of the first of these variance ratios is given by r_{01}^2 and that of the second by r_{02}^2 . This argument may be extended to any number of *uncorrelated component* variables. If the given independent variables do not completely account for the dependent variable, u is added to the right-hand member of the equation to designate the unmeasured causes.

In the typical relationship in education, the independent variables are not components of the dependent variable and are usually correlated. In such cases the dependent variable cannot be precisely expressed as a linear function of the given independent variable and the unmeasured causes. Hence, the equation formed is only an approximate expression of the relationship. In Chapter X, page 324, the regression equation was introduced as the best estimate of a dependent variable that could be made from the given independent variables. For two independent variables

$$\bar{x}_0 = b_{01.2}x_1 + b_{02.1}x_2$$

The expression $b_{01.2}x_1 + b_{02.1}x_2$ gives the best estimates ¹ of x_0 to be made from x_1 and x_2 . Hence it will not be equal to x_0 . This condition makes it necessary to add u to represent the effect of using estimates of the contributions of x_1 and x_2 plus any unmeasured causes. In the following development, u is assumed ² to be uncorrelated with the given independent variables.

¹ The introduction of the concept of a component variable, i.e., one that is contributed completely to the dependent variable, provides a basis for a significant comment upon the multiple regression equation and errors of estimate. If the independent variables are uncorrelated components, the errors of estimate are due to the presence of one or more unmeasured causes, but when the independent variables are not components, this condition serves to increase the errors of estimate. The errors of estimate in predictions are commonly explained to being due to unmeasured causes. It seems likely that a considerable portion of the errors of estimate is due to variable errors of measurement and variable errors of validity in the independent variables. See pages 358 f.

² This assumption is probably only approximated.

$$\begin{aligned}
 \sigma_0^2 &= \frac{\Sigma x_0^2}{N} = \frac{\Sigma (b_{01 \cdot 2} x_1 + b_{02 \cdot 1} x_2 + u)^2}{N} \\
 &= \frac{\Sigma (b_{01 \cdot 2}^2 x_1^2 + b_{02 \cdot 1}^2 x_2^2 + 2b_{01 \cdot 2} b_{02 \cdot 1} x_1 x_2 + u^2 + 2b_{01 \cdot 2} x_1 u + 2b_{02 \cdot 1} x_2 u)}{N} \\
 &= \frac{b_{01 \cdot 2}^2 \Sigma x_1^2}{N} + \frac{b_{02 \cdot 1}^2 \Sigma x_2^2}{N} + \frac{2b_{01 \cdot 2} b_{02 \cdot 1} \Sigma x_1 x_2}{N} + \frac{\Sigma u^2}{N} \\
 &\quad + \frac{2b_{01 \cdot 2} \Sigma x_1 u}{N} + \frac{2b_{02 \cdot 1} \Sigma x_2 u}{N}
 \end{aligned}$$

Since u is assumed to be uncorrelated with either x_1 or x_2 , $\Sigma x_1 u$ and $\Sigma x_2 u$ are zero. Hence the last two fractions in the above equation disappear. The remaining product term may be transformed as follows

$$\begin{aligned}
 \frac{2b_{01 \cdot 2} b_{02 \cdot 1} \Sigma x_1 x_2}{N} &= 2b_{01 \cdot 2} b_{02 \cdot 1} \frac{\Sigma x_1 x_2}{N \sigma_1 \sigma_2} \sigma_1 \sigma_2 \\
 &= 2b_{01 \cdot 2} b_{02 \cdot 1} r_{12} \sigma_1 \sigma_2
 \end{aligned}$$

Hence we have

$$\sigma_0^2 = b_{01 \cdot 2}^2 \sigma_1^2 + b_{02 \cdot 1}^2 \sigma_2^2 + 2b_{01 \cdot 2} \sigma_1 b_{02 \cdot 1} \sigma_2 r_{12} + \sigma_u^2$$

As expressed by this equation, x_1 makes two contributions to σ_0^2 , the variance of x_0 . The "best" estimate of the first is given by $b_{01 \cdot 2}^2 \sigma_1^2$. Call it the direct (individual) contribution. Similar statements may be made with reference to x_2 . The "best" estimates of the other contributions of the two variables are combined in the product term which may be thought of as their joint (indirect) contribution. The remaining term σ_u^2 represents the contributions of the unmeasured causes plus the attenuating effect of using the right-hand member of the regression equation in expressing the functional relationship. The per cents of the total variance may be obtained by dividing both sides of the equation by σ_0^2 .

$$1 = b_{01 \cdot 2}^2 \frac{\sigma_1^2}{\sigma_0^2} + b_{02 \cdot 1}^2 \frac{\sigma_2^2}{\sigma_0^2} + 2b_{01 \cdot 2} \frac{\sigma_1}{\sigma_0} b_{02 \cdot 1} \frac{\sigma_2}{\sigma_0} r_{12} + \frac{\sigma_u^2}{\sigma_0^2}$$

The first term of the right-hand member is the square of the corresponding beta coefficient ¹ of the multiple regression equation and might be written $\beta_{01 \cdot 2}^2$. Similarly, the third term

¹ See page 325.

might be written $2r_{12}\beta_{01\cdot2}\beta_{02\cdot1}$. Several writers have written the equation as follows:

$$\begin{aligned} 1 &= d_{01\cdot2} + d_{02\cdot1} + 2p_{01}p_{02}r_{12} + d_{0u} \\ &= d_{01\cdot2} + d_{02\cdot1} + d_{012} + d_{0u} \end{aligned}$$

The term $d_{01\cdot2}$ is read the coefficient of direct determination of x_1 with respect to x_0 . The symbol p_{01} is read the path coefficient¹ connecting x_0 and x_1 . The term d_{012} is read the coefficient of joint (indirect) determination of x_1 and x_2 with respect to x_0 . The values of the first three terms of the right-hand member of the above equations can be calculated from the data. The value of d_{0u} is obtained by subtracting the sum of these terms from 1.00.

This development, which may be extended to n independent variables, affords a means of interpreting the coefficients of the multiple regression equation in terms of the contributions of the independent variables to the variance of the dependent variable. This interpretation is most conveniently made when the multiple regression equation is expressed in terms of beta coefficients or coefficients of determination. For n independent variables the equation in terms of the latter would be

¹ The term "path coefficient" was proposed by Sewall Wright in developing a different technique for dealing with this problem, but it has been shown that path coefficients are merely the beta coefficients of the regression equation and the expression $b_{01\cdot2}\frac{\sigma_1}{\sigma_0}$ is equal to p_{01} . Since the regression equation is an established statistical procedure, the development given here seems preferable. Wright's work, however, is significant since it provides a basis for utilizing the regression equation in studying functional relationships.

Wright, Sewall. "Correlation and Causation," *Journal of Agricultural Research*, 20: 557-85, January, 1921.

Wright, Sewall. "The Theory of Path Coefficients," *Genetics*, 8: 238-55, May, 1923.

The reader who consults these references to Wright's work will note that the present writers have made slight changes in his symbolism in order to provide a more consistent system.

For proof of the identity of path coefficients and beta coefficients, see

Kelly, E. L. "The Relationship between the Techniques of Partial Correlation and Path Coefficients," *Journal of Educational Psychology*, 20: 119-24, February, 1929.

Dunlap, J. W., and Cureton, E. E. "On the Analysis of Causation," *Journal of Educational Psychology*, 21: 673-75, December, 1930.

$$1 = d_{01 \cdot 23 \dots n} + d_{02 \cdot 13 \dots n} + \dots + d_{0n \cdot 12 \dots (n-1)} \\ + d_{0\bar{1}2 \cdot 34 \dots n} + \dots + d_{0\bar{i}j \dots () \dots () \dots n} + d_{0u}$$

In the next to the last term, i takes any value from 1 to n , and j takes any value from 1 to n other than i . There will be ${}_nC_2$ such terms, i.e., terms for all possible pairs of the independent variables. In this equation for the general case the coefficients of joint determination may be expressed in terms of the coefficient of correlation between the two independent variables and the path coefficients connecting them with the dependent variable. For example,

$$d_{0\bar{1}2 \cdot 34 \dots n} = 2r_{12}p_{01 \cdot 23 \dots n}p_{02 \cdot 13 \dots n}$$

The square of the coefficient of multiple correlation $R_{0 \cdot 123 \dots n}$ is equal to the sum of the coefficients of determination involving the independent variables $x_1, x_2, x_3 \dots x_n$. In other words, $R_{0 \cdot 123 \dots n}^2 = 1 - d_{0u}$. This relationship provides a means for checking the calculations made in computing the several coefficients of determination. It should be noted, however, that when $R_{0 \cdot 123 \dots n}^2$ is used as a measure of the total proportion of the variance of x_0 which is due to $x_1, x_2, x_3 \dots x_n$, it is subject to the limitation noted on page 392. Only, when $x_1, x_2, x_3 \dots x_n$ are uncorrelated components, does $R_{0 \cdot 123 \dots n}^2$ measure the proportion of the variance of x_0 due to these factors with precision. $R_{0 \cdot 123 \dots n}^2$ is called the coefficient of multiple determination.

Interpretation of regression coefficients. The preceding argument suggests a plan for interpreting coefficients of the multiple regression equation as measures of the contributions of the independent variables to the dependent variable of a causal relationship. If the regression equation is not expressed in terms of beta coefficients, the first step is to compute them by means of the relationship given on page 325. Second, compute the squares of the beta coefficients and the products of the type

$$2r_{12}\beta_{01 \cdot 23 \dots n}\beta_{02 \cdot 13 \dots n},$$

These results are coefficients of determination, and hence are

values of variance ratios. The interpretation of coefficients of determination is considered on page 397.

Calculation of path (beta coefficients)—Wright's method. Wright's method of calculating path (beta) coefficients is based on a theorem which may be stated as follows:

Given x_0 , the dependent variable, and $x_1, x_2, x_3 \dots x_n$ independent variables, the coefficient of correlation between x_0 and any one of the independent variables or between any two of the independent variables, is equal to the path coefficient connecting the two variables *plus* the sum of the products of the path coefficients along paths of indirect connection, paths through the dependent variable not being included.¹

The application of this theorem may be explained by employing a diagrammatic representation suggested by Wright and used in a slightly modified form by Burks² and Heilman.³ This plan of representation is shown for three independent variables in Figure 6. Each of the variables and the unknown (remaining component of the dependent variable) are represented by small rectangles, but the reader should bear in mind that neither the size of the rectangles nor their relative position has any significance. The lines connecting the rectangles merely indicate paths of influence. The reader should note that the line connecting two independent variables indicates that the path coefficient may be thought of with respect to either variable as the dependent one. The value of the path coefficient is the same, but if it is desired to recognize the direction of the assumed rela-

¹ For the proof of this theorem see the references to Wright's work on page 393.

² Burks, B. S. "The Relative Influence of Nature and Nurture upon Mental Development; a Comparative Study of Foster Parent-Foster Child Resemblance and True Parent-True Child Resemblance," *The Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1928, pp. 299-301.

³ Heilman, J. D. "The Relative Influence upon Educational Achievement of Some Hereditary and Environmental Factors," *The Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1928, pp. 35-65. For a more extended account, see

Heilman, J. D. "Factors Determining Achievement and Grade Location," *Journal of Genetic Psychology*, 36: 435-56, September, 1929.

tionship, the subscript of the variable considered dependent may be written in first position.

For three independent variables as illustrated in Figure 6, the application of Wright's theorem gives the following equations.¹

$$\begin{aligned} r_{01} &= p_{01} + p_{12}p_{02} + p_{31}p_{03} + p_{12}p_{32}p_{03} + p_{13}p_{23}p_{02} \\ r_{02} &= p_{02} + p_{12}p_{01} + p_{32}p_{03} + p_{32}p_{13}p_{01} + p_{12}p_{31}p_{03} \\ r_{03} &= p_{03} + p_{13}p_{01} + p_{23}p_{02} + p_{13}p_{21}p_{02} + p_{23}p_{12}p_{01} \\ r_{12} &= p_{12} + p_{32}p_{13} \\ r_{13} &= p_{13} + p_{23}p_{12} \\ r_{23} &= p_{23} + p_{13}p_{21} \end{aligned}$$

Usually, we are concerned only with the values of the path coefficients of the paths leading to the dependent variable since

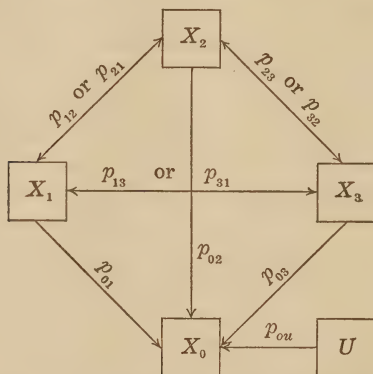


FIG. 6. Path coefficient diagram for three independent variables. p_{01} , p_{02} , and p_{03} are also the beta coefficients $\beta_{01.23}$, $\beta_{02.13}$, and $\beta_{03.12}$.

these are the only ones used in calculating the coefficients of determination of the types $d_{01.234\dots n}$ and $d_{0\bar{1}\bar{2}.34\dots n}$. When the equivalents of r_{12} , r_{13} , and r_{23} are substituted in the equations for r_{01} , r_{02} , and r_{03} , they simplify² to

¹ The path coefficients are written with simplified subscripts. The more elaborate subscripts are useful for indicating the connection with regression coefficients. The subscripts of coefficients of determination may also be simplified.

² Examination of the equations will indicate how they may be extended to any number of independent variables. Furthermore, they may be written without

$$\begin{aligned}
 r_{01} &= p_{01} + r_{12}p_{02} + r_{13}p_{03} \\
 r_{02} &= p_{02} + r_{23}p_{03} + r_{12}p_{01} \\
 r_{03} &= p_{03} + r_{13}p_{01} + r_{23}p_{02}
 \end{aligned}$$

These simultaneous equations may be solved by further substitution, but when there are several independent variables, it is advisable to employ the methods developed for calculating multiple regression equations.¹ After the path (beta) coefficients have been obtained, the calculation of the coefficients of determination is a simple matter.

Interpretation of coefficients of determination. A coefficient of determination represents the value of a variance ratio and should be interpreted as such. A coefficient of direct determination gives a measure of only the direct contributions from an independent variable. A share of one or more of the joint contributions must be added. Wright did not develop a technique for dividing the joint contributions. Heilman divided them in proportion to the direct contributions of the two independent variables involved. In the absence of a better procedure, this method may be employed, but the total coefficients of determination thus obtained should be regarded as only approximations.

After the total coefficients of determination have been obtained, our ignorance concerning what our measures actually represent creates a difficulty. Suppose the independent variables are scores on an intelligence test and scores on a silent reading test and that the total coefficients of determination are found to be .30 and .40. To say that what is measured by the intelligence test contributes 30 per cent of the variance of the dependent variable is not satisfactory because we do not know what the intelligence test measures. Obviously a portion of the complex of abilities it measures is also measured by the silent reading test. More meaningful findings would be coefficients of determination for the factor of the intelligence test scores that

constructing the representation illustrated in Figure 6. They are identical, except for symbolism, with the "normal equations" given on page 332.

¹ See pages 330-32. The method developed by Griffin is economical.

is uncorrelated with the silent reading test scores, the factor common to the two variables, and the factor of the silent reading test scores that is uncorrelated with the intelligence test scores. In explaining the general problem for which he developed the path coefficient technique, Wright refers to such causes which he designates as "remote" but he does not give any technique for identifying and measuring their contributions. Until we are able to identify the remote (elemental) causes and to secure measures of their contributions, an investigator is restricted in interpreting coefficients of determination derived from the coefficients of a multiple regression equation.

Dependability of coefficients of determination derived from the coefficients of the multiple regression equation. In deriving the equation

$$1 = d_{01.2} + d_{02.1} + d_{01\bar{2}} + d_{0u}$$

the multiple regression equation was used. (See page 391.) It gives only the best estimates of x_0 to be obtained from x_1 and x_2 . This condition operates to reduce the calculated value of the coefficients of determination. What happens may be illustrated by using a set of variables built up from counts of coin tosses as follows:

$$\begin{aligned} X_0 &= A_1 + A_2 + A_3 + A_4 + A_5 \\ X_1 &= A_1 + A_2 + A_6 \\ X_2 &= A_1 + A_3 + A_7 \\ X_3 &= A_1 + A_4 + A_8 \end{aligned}$$

The set-up represented by these variables is somewhat similar to that which we have when X_0 represents measures of achievement and X_1 , X_2 , and X_3 are measures of abilities that contribute to this achievement. The component A_1 may be thought of as corresponding to Spearman's "g" factor.¹ The components A_6 , A_7 , and A_8 may be thought of as representing variable errors of measurement and validity. By employing the path

¹ See pages 402-03.

coefficient technique, the following values were obtained for the coefficients of determination:

$$\begin{array}{r}
 d_{01} = .1901 \\
 d_{02} = .1116 \\
 d_{03} = .0400 \\
 d_{012} = .1044 \\
 d_{013} = .0682 \\
 d_{023} = .0536 \\
 \hline
 \text{Total} \quad .5679
 \end{array}$$

Subtracting this total from 1.00, we have $d_{0u} = .4321$. By the definition of X_0 , $u = A_5$. Direct calculation gives $d_{0A_5} = .100$. This means that the calculated values for d_{01} , d_{02} , etc., are too small. Their total should be .90 instead of .5679. The attenuation of the calculated values is due to the use of "best" estimates in forming the equation of relationship, which, in this case would be

$$X_0 = b_{01.23}X_1 + b_{02.13}X_2 + b_{03.12}X_3 + U$$

This condition should be kept in mind when the coefficients of determination are calculated.

Concept of elemental causes. On pages 397-98 the point was made that coefficients of determination for uncorrelated causes would be more meaningful statistics. Such causes may be described as elemental with respect to the group of independent variables considered.¹ Let a_1 , a_2 , a_3 , a_4 , a_5 , and a_6 represent causes contributing to x_0 which are elemental with respect to the given independent variables. Also, let s_0 represent the sum of all contributing factors that are unrelated to the given independent variables and let e_0 represent the chance factor included in the dependent variable. Assuming that x_0 is a linear function of its causes, it may be expressed as

$$x_0 = c_{01}a_1 + c_{02}a_2 + c_{03}a_3 + c_{04}a_4 + c_{05}a_5 + c_{06}a_6 + c_{07}s_0 + c_{08}e_0$$

¹ They are also elemental with respect to x_0 here designated as the dependent variable. When elemental causes are being considered any variable of the group might be designated as dependent.

The independent variables may also be expressed as weighted sums. For example, they might be defined as follows:

$$\begin{array}{llll}
 x_1 = & c_{11}a_1 & + c_{13}a_3 + c_{14}a_4 & + c_{17}s_1 + c_{18}e_1 \\
 x_2 = & c_{22}a_2 & + c_{23}a_3 + c_{24}a_4 + c_{25}a_5 & + c_{27}s_2 + c_{28}e_2 \\
 x_3 = & c_{31}a_1 & + c_{33}a_3 & + c_{35}a_5 + c_{37}s_3 + c_{38}e_3 \\
 x_4 = & c_{41}a_1 & + c_{43}a_3 + c_{44}a_4 & + c_{46}a_6 + c_{47}s_4 + c_{48}e_4 \\
 x_5 = & c_{52}a_2 & + c_{53}a_3 & + c_{56}a_6 + c_{57}s_5 + c_{58}e_5
 \end{array}$$

This analytical pattern is probably typical of those encountered in educational research. However, one cannot assume the absence of an elemental component in any of the independent variables. Hence, the blank spaces must be assumed to be filled until there is convincing evidence that certain c 's (factor loadings) are zero.

If additional independent variables are introduced into a given situation, the number of a 's may be increased, but it is reasonable to assume that eventually all possible a 's will be represented in the pattern. In such a situation, the a 's may be thought of as elemental causes in an absolute sense. It should be noted that the relationship between the dependent variable and the elemental causes is always one of cause and effect. Hence, in studying the contributions from elemental causes, it is not necessary to demonstrate that the independent variables are causally related to the dependent variable.

Measurement of the contributions from elemental causes.

The contribution of a_1 to x_0 is given by the variance ratio $\frac{c_{01}^2 \sigma_{a_1}^2}{\sigma_0^2}$ and the contribution of the other a 's by corresponding variance ratios. If the a 's, s_0 , and e_0 , are expressed in terms of standard units, $\sigma_{a_1}^2 = \sigma_{a_2}^2 = \sigma_{a_3}^2 \dots = \sigma_{s_0}^2 = \sigma_{e_0}^2 = 1$. Furthermore, the c 's may be chosen so that $\sigma_0^2 = 1$. Hence, the variance ratios reduce to squares of the c 's and we have

$$1 = c_{01}^2 + c_{02}^2 + c_{03}^2 + c_{04}^2 + c_{05}^2 + c_{06}^2 + c_{07}^2 + c_{08}^2$$

The problem of measuring the contributions of the elemental components to the dependent variable is one determining the squares of the c 's in the equation of the dependent variable.

Similar equations may be written for each of the independent variables. A reliability coefficient is equal to 1.00 minus the square of the factor loading of the corresponding chance factor. For example,

$$r_{00} = 1.00 - c_{08}^2$$

The correlation between two x 's is expressible in terms of c 's.¹ For example, in the situation illustrated here

$$r_{01} = c_{01}c_{11} + c_{02}c_{12} + c_{03}c_{13} + c_{04}c_{14} + c_{05}c_{15} + c_{06}c_{16}$$

When the number of equations thus formed is equal to the number of c 's, it is theoretically possible to determine their values. However, one does not know in advance how many c 's are involved and hence the determination of their values cannot be accomplished by writing and solving a set of simultaneous quadratic equations. The problem is that of obtaining a unique determination of the values of the c 's from the intercorrelations. Kelley² has proposed a method of successive approximations based upon least squares, but Holzinger³ has shown that several different solutions may be fitted to Kelley's data. Recently Thurstone⁴ has developed a technique which gives, in the case of certain factor patterns, a unique determination of the c 's (factor loadings) for the factors common to two or more

¹ Provided the x 's and their factors are in terms of standard units.

² Kelley, T. L. *Crossroads in the Mind of Man*. Stanford University, California: Stanford University Press, 1928, pp. 122 f.

Another reference dealing with the same problem is Hotelling, Harold. "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, 24: 417-41, 498-520, September, October, 1933.

³ Holzinger, Karl J., and Swineford, Frances. "Uniqueness of Factor Patterns," *Journal of Educational Psychology*, 23: 247-58, March, 1932.

⁴ Thurstone, L. L. *The Theory of Multiple Factors*. Ann Arbor, Michigan: Edwards Brothers, Inc., June, 1932. 65 pp.

Thurstone, L. L. *A Simplified Multiple Factor Method and an Outline of the Computations*. Ann Arbor, Michigan: Edwards Brothers, 1933. 26 pp.

Thurstone, L. L. "Vectors of Mind," *Psychological Review*, 41: 1-32, January, 1934.

Thurstone, L. L. "Unitary Traits," *Journal of General Psychology*, 11: 126-32, July, 1934.

A more comprehensive reference is Thurstone, L. L. *The Vectors of Mind*. Chicago, University of Chicago Press, 1935. 266 pp.

of the variables. The c for the specific factor of a variable may be obtained by subtracting from unity the sum of the squares of the c 's of the common factors included in the variable and the square of the factor loading of the chance factor obtained by means of the coefficient of reliability.

Determining the presence of a component common to four or more variables. The question of the conditions under which a group of intercorrelated variables may be considered to involve a common component attracted the attention of Spearman thirty years ago.¹ He noted that in some cases the coefficients of intercorrelation between variables tended to fit a mathematical formula, called the *tetrad equation*, which in the case of four variables may be written

$$\frac{r_{12}}{r_{13}} = \frac{r_{24}}{r_{34}}$$

This may be changed to the form

$$r_{12}r_{34} - r_{13}r_{24} = 0$$

The left-hand member of the equation, designated by the symbol t_{1234} , is called a *tetrad difference*. For four variables, there are three tetrad differences:

$$\begin{aligned} t_{1234} &= r_{12}r_{34} - r_{13}r_{24} \\ t_{1243} &= r_{12}r_{34} - r_{14}r_{23} \\ t_{1342} &= r_{13}r_{24} - r_{14}r_{23} \end{aligned}$$

Spearman² formulated the theorem that when the tetrad

¹ Spearman, C. "General Intelligence Objectively Determined and Measured," *American Journal of Psychology*, 15: 201-92, 1904.

² For the first formulation, see Spearman, *op. cit.* The formulation given in his *The Abilities of Man*, pp. 74-75 is essentially the same as the earlier one. See also Spearman, C. "What the Theory of Factors Is Not," *Journal of Educational Psychology*, 22: 112-17, February, 1931.

The reader who is interested in the mathematical proof of the theorem should consult the Appendix of Spearman's *The Abilities of Man* or Line, W., and Hedman, H. B. "A Simplified Statement of the Two-Factor Theory," *Journal of Educational Psychology*, 24: 195-220, March, 1933. This reference gives a bibliography for a more extensive study. A bibliography up to 1928 is given by Walker, H. M. *Studies in Statistical Method*. Baltimore: The Williams and Wilkins Company, 1929, Chapter VI. For a number of theorems, including this one, see Kelley, T. L. *Crossroads in the Mind of Man*. Stanford University,

differences for a group of variables are equal to zero, we have a necessary and sufficient condition for the conclusion that the several abilities represented by the measures include a common factor and that each ability is made up of this common factor (called " g ") and a specific factor ¹ (called " s ").

When a group of variables have a common component and no group factors, the tetrad differences calculated from an actual table of correlation coefficients are likely not to be precisely zero, especially if the number of cases is not large.² Any condition that influences the value of the calculated coefficient of correlation ³ is likely to influence the tetrad differences in which it appears. If the data form a random sample, and the investigator desires to generalize to the larger population or universe, the effect of chance must be considered. Hence, the question arises in regard to the deviation of the tetrad differences from zero that may exist when the variables actually consist of only a general factor and a specific factor. The probable effect of chance may be calculated ⁴ as in the case of other statistics.

California Stanford University Press, 1928, Chapter III. Proposition 10 is Spearman's theorem.

¹ A specific (unique) factor is uncorrelated with the common factor and with the specific factors in the other variables.

² Spearman, C. "Disturbers of Tetrad Differences: Scales," *Journal of Educational Psychology*, 21: 559-73, November, 1930.

For illustrations of distributions of tetrad differences, see Rogers, K. H. " 'Intelligence' and 'Perseveration' Related to School Achievement," *Journal of Experimental Education*, 2: 35-43, September, 1933.

Line, W., and Kaplan, E. "Variation in I.Q., at the Preschool Level," *Journal of Experimental Education*, 2: 95-100, December, 1933.

Rogers, K. H. "Perseveration in a Group of Subnormal Children," *Journal of Experimental Education*, 2: 301-09, March, 1934.

³ See pages 101 f., 151, and 154.

⁴ Several formulae have been developed.

Spearman, C. *The Abilities of Man*. New York: The Macmillan Company, 1927, Appendix, p. xi. (Formula 16a.) The proof of this formula has been promised but not published.

Kelley, T. L. *Crossroads in the Mind of Man*. Stanford University, California: Stanford University Press, 1928, p. 49.

Moul, M., and Pearson, K. "The Mathematics of Intelligence. I. The Sampling Errors in the Theory of a Generalized Factor," *Biometrika*, 19: 246-91, 1927.

Wishart, John. "Sampling Errors in the Theory of Two Factors," *British Journal of Psychology*, 19: 181-87, June, 1928.

An empirical test of the formulae has been reported by Garrett, H. E. "The

Thurstone has stated as a general theorem that the necessary number of uncorrelated (orthogonal) components (factors) in a group of variables is equal to the "rank" of the table of their intercorrelations considered as a determinant.¹ The "rank" of a determinant is the order of the highest order of minors, all of which are not "equal to zero."² This means that if all of the fourth-order minors are equal to zero but all of the third-order minors are not equal to zero, then the table of intercorrelations may be explained by the presence of three uncorrelated components. Since the tetrad differences are merely the second-order minors, Spearman's theorem is a special case of Thurstone's theorem.³

Contributions of factors common to three or more independent variables. When it is apparent that a group of variables includes a common factor and no group factor, the correlation of the common factor with any one of the variables may be calculated by the formula ⁴

$$r_{ag} = \frac{\sqrt{A^2 - A'}}{\sqrt{T' - 2A}}$$

Sampling Distribution of the Tetrad Equation," *Journal of Educational Psychology*, 24: 536-42, October, 1933.

¹ Thurstone, L. L. *The Theory of Multiple Factors*. Ann Arbor, Michigan: Edwards Brothers, Inc., 1932, p. 20.

² "Equal to zero" is to be interpreted in the sense that a group of tetrad equations are considered equal to zero.

³ The reader interested in factor analysis will find the following references helpful:

Chant, S. N. F. "Multiple Factor Analysis and Psychological Concepts," *Journal of Educational Psychology*, 26: 263-72, April, 1935.

Russell, O. R. "Some Observations on Multiple-Factor Analysis," *Journal of Educational Psychology*, 26: 284-85, April, 1935.

Thomson, G. H. "The Definition and Measurement of "g" (General Intelligence)," *Journal of Educational Psychology*, 26: 241-62, April, 1935.

⁴ Spearman, C. *The Abilities of Man*. New York: The Macmillan Company, 1927, Appendix, p. xvi.

For applications of this formula, see Holzinger, K. J. "Thorndike's C.A.V.D. Is Full of g," *Journal of Educational Psychology*, 22: 161-66, March, 1931.

Cairns, G. J. "An Analytical Study of Mathematical Abilities," *The Catholic University of America, Educational Research Monographs*, Vol. 6, No. 3. Washington, D. C.: The Catholic University Press, 1931. 104 pp.

Cairns applied the tetrad technique and certain supplementary procedures to data obtained by administering eighteen tests.

A is the sum of the intercorrelations between Test a and every other test of the group, A' is the sum of the squares of these correlations, and T is the total of all intercorrelations. The coefficient r_{ag} is a measure of the extent to which the scores of a test are "saturated" with " g " and has been called the "coefficient of saturation." These correlations are equivalent¹ to the factor loadings of the first factor resulting from analysis of a correlation matrix by means of the "centroid" method developed by Thurstone.² Continuation of this analysis results in the determination of the factor loadings³ of any other common factors which may be present among the variables, but the factor loadings thus secured must be subjected to further mathematical treatment if a "unique" solution is to be obtained.

Applications of correlation analysis. The techniques described in the preceding pages have a number of applications. Multiple regression equations with beta coefficients are used to determine the appropriate weights to be assigned to the subtests of a battery in computing the total score.⁴ Since the beta coefficients are equal to the path coefficients connecting the subtests represented by $X_1, X_2 \dots X_n$ to the criterion X_0 , the coefficients of determination for the independent (direct) contributions to the criterion and for the joint contributions of the pairs of the sub-tests can easily be obtained. The sum of the coefficients of determination for a given group of such tests represents the statistical validity⁵ of the battery as determined

¹ If all of the tetrads vanish, the two methods give identical results.

² Thurstone, L. L. *A Simplified Multiple Factor Method and an Outline of the Computations*. Ann Arbor, Michigan: Edwards Brothers, 1933. 26 pp.

³ Factor loadings are the c 's previously referred to. They are the correlations of the variables with the elemental or primary factors. When these factors are orthogonal, or uncorrelated, the factor loadings are the beta coefficients of standard score regression equations in which the elemental, or primary, factors are the independent variables. The squares of the factor loadings measure the proportions of the variance of a given variable due to the primary factors.

⁴ See Kelley, T. L. *Interpretation of Educational Measurements*. Yonkers-on-Hudson: World Book Company, 1927, pp. 212-13.

⁵ For definitions of reliability and validity in terms of variance, see Cureton, E. E. "Errors of Measurement and Correlation," *Archives of Psychology*, No. 125, 1931, pp. 8-13.

Cureton, E. E. "Validation against a Fallible Criterion," *Journal of Experimental Education*, 1: 258-63, March, 1933.

by the criterion and the separate coefficients of determination will picture the contributions of the several sub-tests. A relatively high value of a coefficient of determination of the type $d_{01.23 \dots n}$ indicates a sub-test that contributes significantly to the measurement of the desired ability or trait. A relatively low coefficient of determination of the type $d_{0\overline{12.34 \dots n}}$ indicates relatively little overlapping in function by the sub-tests represented by X_1 and X_2 .

Factor analysis appears to offer a means of isolating and identifying the human traits and abilities that we wish to measure. It is not unreasonable that eventually a group of elemental abilities and traits may be identified and described which will bear somewhat the same relation to achievement in the various fields that chemical elements bear to chemical compounds. Such determinations will contribute materially to the construction of educational tests. Factor analysis also appears to have possibilities in connection with problems having to do with cause and effect relationships.

Variance ratios, factor loadings, and some of the other terms introduced are relatively new and an investigator employing correlation analysis should make certain that he understands the interpretation of the results yielded by these techniques. On page 372 it was pointed out that the contribution from a cause is to the variability (variance) and not to the raw measures of the variable. This means that the results obtained by means of correlation analysis are for a population of certain specifications. Hence, it is important that the collection of data be wisely planned. Otherwise the results obtained may not be for the population in which one is interested.

ILLUSTRATIONS OF THE APPLICATION OF THE TECHNIQUES OF CORRELATION ANALYSIS

The references of this list were selected to illustrate the application of the techniques of correlation analysis. The reader, however, should not consider them model studies. In some cases the technique employed probably failed to accomplish what was desired.

ADAMS, H. F. "A Non-Intellectual General Factor," *Journal of Educational Psychology*, 23: 173-78, March, 1932.

This study is presented by its author as evidence that tetrad technique is applicable to correlations relating to other phenomena than mental abilities. The data used were "scores made by the most outstanding amateur and professional golfers in the three major tournaments held in this country during the past eleven years." Length of holes, as a common factor, is shown to be similar to Spearman's "*g*."

BALDWIN, B. T. "The Relation between Mental and Physical Growth," *Journal of Educational Psychology*, 13: 193-203, April, 1922.

Zero order coefficients are reported for physical and mental traits of 49 girls and partial correlation was employed to eliminate chronological age. A range of several years was eliminated by this technique, but the regression of height on age appears markedly curvilinear as shown by "Growth Curves in Height." On the other hand, for the age range studied, the mental growth curves are apparently linear.

BURKS, B. S. "The Relative Influence of Nature and Nurture upon Mental Development; a Comparative Study of Foster Parent-Foster Child Resemblance and True Parent-True Child Resemblance," *The Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1928, pp. 219-316.

On pages 299-304 is reported a path coefficient study in which child intelligence is the dependent variable and measures of parental intelligence and of environment are the independent variables. In place of rationing off the joint influence of the independent variables according to the weights of the direct influences, Burks computed a *coefficient of part determination* in an effort to show the total contribution of parental intelligence to the variance of child intelligence, where only those aspects of environment are held constant which are independent of parental intelligence.

CAIRNS, G. J. "An Analytical Study of Mathematical Abilities," *The Catholic University of America, Educational Research Monographs*, Vol. 6, No. 3. Washington: The Catholic University Press, 1931. 104 pp.

The tetrad technique was applied in this study to intercorrelations between scores on eighteen tests. In an effort to locate group factors, "*g*" was eliminated from these intercorrelations. Where the net coefficients appeared of significant magnitude, the reference variable technique was applied on the hypothesis that tetrad differences significantly different from zero would locate a group factor in the two variables other than the reference ones.

COLLINS, J. E. "The Intelligence of School Children and Paternal Occupation," *Journal of Educational Research*, 17: 157-69, March, 1928.

The problem of this study was to determine the degree of relationship between intelligence of children and the occupational status of the father. No coefficients of correlation are reported, but the nature of the relationship is probably more adequately indicated in tables and graphs. For example, the interquartile ranges of intelligence quotients of groups of children, classified according to the following occupational groups: agriculture, unskilled labor, skilled labor, foreman, trade, clerical, managerial, and professional, indicates a concomitant increase of intelligence with variation of occupation from unskilled labor to the managerial type.

DENWORTH, K. M. "The Effect of Length of School Attendance upon Mental and Educational Ages," *The Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1928, pp. 67-91.

Coefficients of partial correlation and standard score regression equations are reported in this study. The author recognizes the limitations of the partial correlation technique and restricts her interpretation of the regression equations to their prognostic significance.

HERRIOTT, M. E. "Attitudes as Factors of Scholastic Success," *University of Illinois Bulletin*, Vol. 27, No. 2, *Bureau of Educational Research Bulletin*, No. 27. Urbana: University of Illinois, 1929. 72 pp.

The groups studied consisted of 260 students of educational psychology and 113 students of technique of teaching. The variables measured included intelligence, reading ability, study habits and attitudes. The partial correlation technique was employed up to the calculation of *tenth-order* partials.

HOLZINGER, K. J. "Thorndike's C.A.V.D. Is Full of *g*," *Journal of Educational Psychology*, 22: 161-66, March, 1931.

In this study, the tetrad technique is applied to intercorrelations between C.A.V.D., Otis Self-Administering Test, Terman Group Test, and Stanford Binet. The correlations of each with *g* are reported respectively .960, .921, .960, and .817. Hence, C.A.V.D. is "full of *g*."

KELLEY, T. L. *Crossroads in the Mind of Man*. Stanford University, California: Stanford University Press, 1928. 238 pp.

In this comprehensive study of factor theories, Kelley presents his criticisms of Spearman's work, states a number of propositions with respect to general and specific factors in terms of varying numbers of variables, develops techniques and applies them to data collected from kindergarten, third-grade, and seventh-grade children.

KELLEY, T. L. *The Influence of Nurture upon Native Differences*. New York: The Macmillan Company, 1926. 49 pp.

After carefully defining the terms "nature" and "nurture," Kelley develops certain techniques for use in differentiating between nature and nurture factors and in abstracting the nurture factor. He applies his techniques to data obtained from eight-, eleven-, and fourteen-year-old children.

LINE, W., ROGERS, K. H., and KAPLAN, E. "Factor Analysis Techniques Applied to Public-School Problems," *Journal of Educational Psychology*, 25: 58-65, January, 1934.

This study includes applications of both Spearman's and Thurstone's techniques. It is stated that the latter "Made possible a more exhaustive factorial analysis, although the significance of the subsidiary factors cannot yet be stated."

NANNINGA, S. P. "Costs and Offerings of California High Schools in Relation to Size," *Journal of Educational Research*, 24: 356-64, December, 1931.

In this study curvilinear relationships were discovered which necessitated the calculation of ratios of correlation and the use of curve-fitting by the method of least squares.

SLOCOMBE, C. S. "Of Mental Testing—A Pragmatic Theory," *Journal of Educational Psychology*, 19: 1-24, January, 1928.

An excellent discussion of the two-factor theory. Worthy of careful study by the interested student.

SNEDECOR, G. W. "Calculation and Interpretation of Analysis of Variance and Covariance," *Monograph Number One, Division of Industrial Science*, Iowa State College. Ames, Iowa: Collegiate Press, Inc., 1934. 96 pp.

This reference deals with the techniques to be employed in analyzing the total variance of a set of measurements to show the contributions of various factors of classification, or of heterogeneity. These techniques represent an extension of the work of R. A. Fisher. (See page 252.) While illustrations of the techniques are given from the fields of agriculture and biology, they are worthy of application to educational problems. For example, the problem of determining whether freshmen classes over a period of years represent a homogeneous population with respect to intelligence, or other traits, may be attacked through the use of techniques described here. Another problem, for which the techniques are applicable, is that of determining the extent to which the correlation between certain traits is an internal characteristic of several groups, or an inter-group relationship.

SPEARMAN, C. E. *The Abilities of Man*. New York: The Macmillan Company, 1927. 415 pp.

In this comprehensive text, Spearman discusses several theories, including his own, concerning the nature of intelligence and other human abilities. The research of over twenty years on the two-factor theory is summarized in this volume. The appendix is an excellent source of information with respect to the statistical techniques employed.

SYMONDS, P. M. "The Effect of Attendance at Chinese Language Schools on Ability with the English Language," *Journal of Applied Psychology*, 8: 411-23, December, 1924.

Bi-serial coefficients of correlation were calculated in this study in an effort to show the relationship between attendance and non-attendance at Chinese language schools (in Hawaii) and ability in English quantitatively measured. Age and intelligence were held constant by means of the partial correlation technique. The application of the bi-serial r technique seems appropriate. A continuous variable may be assumed to underly attendance and non-attendance at a Chinese language school. Elimination of the variable chronological age by partial correlation likewise seems appropriate, but it is probable that intelligence "as measured" does not contribute completely. Furthermore, the application of partial correlation to coefficients, some of which are bi-serial coefficients, seems questionable. The bi-serial r is only approximately equivalent to the Pearson product-moment r .

CHAPTER XII

DETERMINING WHAT SHOULD BE

General character of the problem. The problems dealt with in the preceding chapters may be classified under three heads: (1) What has been? (2) What is? (3) What will be? We may also ask what should be, or what is desirable. Problems asking what has been are known as historical. We have no established designation for those under the second and third captions, but in the absence of a more appropriate term they may be referred to as "scientific." Questions that ask what should be may be designated as "problems of purposes." A significant characteristic of such problems is suggested by the nature of the process of answering them. With reference to pupil failures at a particular grade level, we may ask what per cent of pupils fail or what will be the effect of a specified change in the plan of school organization, or in the method of instruction, or in the curriculum. The answers to such questions may be derived from objective data and when obtained are regarded as facts. A statement of the per cent of pupils that should fail represents a judgment.

The statement that the determination of what should be involves judgment is not intended to imply that the answer to such a question must be a mere opinion. In making a judgment, reflective thinking is involved. The problem is defined, data are collected, and consideration is given to the probable consequences of adopting different possible courses of action in an effort to determine which one is most likely to result in the attainment of an end judged to be desirable. For example, if the problem is to determine whether the assignment method or the project method should be employed in a particular school situa-

tion,¹ a person should inquire into the various aspects of the particular situation—traits of the children, available equipment, traditions of the school, attitude of the community, training and preferences of the teachers, competence of the supervision, etc.—and into the probable effectiveness of the two methods including not only the immediate achievements of the pupils but also their attitudes and preparation for future work. With these considerations in mind he will then decide what appears desirable.

The range of problems of purposes. The field of curriculum construction furnishes a large group of problems of purposes, but questions that ask what should be, arise in other divisions of the field of education. The following are illustrative of such problems.

1. What means should be used in securing publicity for a school building program?
2. What is the value of the daily assembly in high school?
3. How should high school athletic funds be administered?
4. Should applicants for admission to college be given an entrance examination?
5. Should the college resort to careful selection of students for training as teachers?
6. What would constitute a satisfactory minimum program for equalization of educational opportunity for which the state should assume responsibility?
7. What should be the limits of state control of education?
8. What should be the physical training and health program in the elementary school?
9. What should be the relations between superintendents and business managers?
10. Should foreign language credits be required for college entrance?
11. To what extent should pupils participate in school administration?
12. What should be the minimum preparation of junior college instructors?
13. What types of bonds should be issued in financing the construction of school buildings?

¹ The reader should note that the question implied here is not the same as, "What is the relative effectiveness of the assignment method and the project method?"

Further consideration of problems of purposes. Many problems that ask what should be involve two types of questions: (1) a question of objectives and (2) a question of means. For example, consider the first question in the preceding list: "What means should be used in securing publicity for a school building program?" In attempting to answer this question, it is necessary to consider first what objective is to be attained as a result of the publicity. Is approval of the proposed building program by the voters of the district desired? Or, is the objective to secure the decision that will be most beneficial to the educational interests of the community? After the objective has been decided upon, there remains the question of means—that is, how may the desired end be most efficiently attained? In some problems the subordinate questions are somewhat different. For example, the question concerning the value of the daily assembly in the high school may be analyzed as follows: (1) What is the effect of various types of daily assembly in the high school? (2) What is contributed by this effect to the objectives of the school? The second question implies the problems of determining the objectives of the school.

The determination of objectives frequently raises many subordinate questions. For example, in determining the objectives of a high school in a rural area, such questions as the following should receive consideration. Should the pupils be educated so that they will be better adapted to rural living or should they be educated so that they will be better adapted to living in an urban area, or so that they will be better adapted to living in an environment that may be either urban or rural? If the purpose is adaptation to living in rural areas, what type of rural life should be postulated? What portion of the total preparation for rural living shall be made the responsibility of the high school? In a particular situation it will be necessary to consider the resources of the community, its probable future development, the interests and capacities of the children to be educated, the training and experience of available teachers, and other practical aspects of the situation.

How problems of purposes are dealt with. The determination of objectives or purposes requires procedures commonly designated as *philosophical*. Objective data may be collected and the results of scientific studies may be utilized, but the answer is essentially a judgment which represents what is considered to be most desirable. In contrast, the procedures employed in dealing with problems that ask what is or what will be, are designated as *scientific*. The reader, however, should not be misled by this use of the terms philosophical and scientific. They are introduced as a means of convenience in discussing certain topics and not as designating procedures that have nothing in common. Questions of means suggest the experimental method, but in practice it is seldom feasible to deal with them by this procedure and hence the philosophical method is applied to them also.

The philosophical method. In its general outline the method of philosophy is fundamentally the same as that described as the method of research in Chapter I. The philosopher defines his problem, collects data, formulates hypotheses and verifies them. In contrast to the scientific investigator, the philosopher deals with a wider range of data. He does not limit himself to those that may be collected by means of the techniques described in Chapter III. He includes the results of his own observation and facts and principles from related fields. He may seek the experiences of other persons and their beliefs. In working with his data, his method is seldom statistical. In verifying his hypotheses he is more concerned with their implications and their relations to experience and to general principles.¹

¹ For further discussion of the methods of philosophy in contrast to the methods of science, see

Lepley, Ray. "Dependability in Philosophy of Education," *Teachers College, Columbia University Contributions to Education*, No. 461. New York: Bureau of Publications, Teachers College, Columbia University, 1931, pp. 1-15.

Buckingham, B. R. "The Philosophy and Organization of Research," *School and Society*, 29: 755-64, June 15, 1929.

Clugston, H. A., and Davis, R. A. "Is a Scientific Method Possible for Philosophical Research?" *Educational Administration and Supervision*, 12: 293-99, April, 1930.

Clugston, H. A., and Davis, R. A. "Suggested Criteria for the Philosophical

The method of philosophy cannot be described with the definiteness that has been possible in the preceding chapters. Neither is it possible to subject a procedure of a subjective, or non-overt, nature to the same sort of scrutiny that is possible for such a technique as that of securing equivalent groups by pairing pupils on the basis of significant characteristics. We may, however, consider the principal phases of the method of philosophy as it is applied to educational problems that ask what should be.

In defining his problem, the philosopher seeks to identify the fundamental questions involved and to formulate the assumptions upon which the solution of the problem is to be based. In formulating these assumptions he may be guided by his own experiences, but frequently he consults various sources.¹ The validity of the assumptions formulated depends on the availability and dependability of existing human knowledge relevant

Method of Research in Education," *Educational Administration and Supervision*, 16: 575-80, November, 1930.

Hullfish, H. G. "The Relation of Philosophy and Science in Education," *Journal of Educational Research*, 20: 159-65, October, 1929.

Kelley, T. L. "A Defense of Science in Education," *The Harvard Teachers Record*, 1: 123-30, November, 1931.

Kelley, T. L. "The Scientific versus the Philosophic Approach to the Novel Problem," *Science*, 71: 295-302, March 21, 1930.

Kilpatrick, W. H. "A Defense of Philosophy in Education," *The Harvard Teachers Record*, 1: 117-22, November, 1931.

Kilpatrick, W. H. "The Relation of Philosophy to Scientific Research," *Journal of Educational Research*, 24: 97-114, September, 1931.

¹ Freeman, Reisner, and Peters have contended that psychology, history of education, and educational sociology are the sources of fundamental assumptions. The writers feel that *any* relevant body of tested knowledge should be consulted.

Freeman, F. N. "Psychology as the Source of Fundamental Assumptions in Education," *Educational Administration and Supervision*, 14: 371-77, September, 1928.

Reisner, E. H. "The History of Education as a Source of Fundamental Assumptions in Education," *Educational Administration and Supervision*, 14: 378-84, September, 1928.

Peters, C. C. "Educational Sociology as a Source of Fundamental Assumptions in Education," *Educational Administration and Supervision*, 14: 385-92, September, 1928.

The following article should be read along with those referred to above:

Bode, B. H. "Where Does One Go for Fundamental Assumptions in Education?" *Educational Administration and Supervision*, 14: 361-70, September, 1928.

to the problem, and the extent to which this knowledge is intelligently brought to bear on the problem. It is also conditioned by the intelligence of the investigator, his beliefs and his prejudices and by other factors that influence the quality of his thinking.

Interpretation of data in dealing with problems of purposes is accomplished by thinking of hypotheses and then checking them against all available criteria until the most satisfactory one is found. The formulation of hypotheses depends upon a person's sensitiveness to the meaning of his data, the breadth of his acquaintance with the field of his problem, his type of mind, and his fundamental philosophy of life. It is to be expected that some hypotheses will be unsatisfactory. The important considerations are that each hypothesis be carefully checked against the available criteria and that eventually an acceptable one be found. This tested hypothesis is the conclusion, the answer to the problem, but the critical worker will regard it as a judgment subject to modification in the light of any new data that may come to his attention.

A person employing the philosophical method critically examines his procedure and his tentative conclusions. He raises such questions as: "Have I been logical in my thinking about the matter?" "Have I considered all that is relevant to the situation?" "Have I succeeded in being open-minded in arriving at a decision?" "Have I suspended judgment long enough to arrive at a decision which is a reasonably safe basis for action?" Another test of the defensibility of a decision with respect to what should be, is its fruitfulness in further thinking about the matter. Finally, the consequences of acting in accordance with the decision provide the ultimate test of the defensibility of the decision and the basis for the formulation of modifications of the decision.

The contribution of objective techniques to the solution of problems of purposes. Although the determination of what should be involves judgment and the total procedure must be described as subjective, significant contributions may be obtained

from objective techniques. Facts and principles are needed for making intelligent judgments. For example, when a superintendent is considering the desirability of increasing the number of teachers in order to reduce the size of classes, he will wish to know the effect of size of class upon pupil achievement. He probably will desire to know the practices in other school systems of similar size and resources. Objective techniques may be useful in securing the answer to a problem of purposes after an assumption or hypothesis is made with reference to the basis from which the answer is to be derived. For example, if the hypothesis is made that pupils should learn to spell the words that adults use in writing, objective techniques may be employed to determine the words that adults use and the frequency of their use.

Philosophical procedures in dealing with problems of science.

At various places in the preceding chapters the attention of the reader has been directed to assumptions that are basic to the techniques employed in dealing with questions which ask what is or what will be or that are introduced in interpreting the findings of such studies. Identifying these assumptions and arriving at an estimate of their validity in particular cases calls for the methods of philosophy. Hence, in practice the distinction between scientific research and "philosophical research" breaks down.¹ A person dealing with problems of purposes frequently needs the assistance of objective techniques and the worker dealing with problems of science should at times engage in philosophizing. The need for philosophizing is especially urgent in experimental research: When a question is asked concerning the relative effectiveness of two procedures or practices, it is generally assumed that an experimental solution of the problem is possible, but, as pointed out on page 291, this assumption may

¹ Symonds contends that it is inappropriate to recognize philosophical research as a type and his position is defensible if the meaning of research is restricted to that of a scientific investigation. However, the use of the term may be justified as a convenience. Symonds, P. M. "A Course in the Technique of Educational Research," *Teachers College Record*, 29: 29-30, October, 1927.

not appear to be defensible when the problem is adequately defined.

Dependability in science versus dependability in philosophy. There is a widespread tendency to consider objectivity of data the criterion of the dependability of a conclusion. If the data employed are highly objective, many persons classify the conclusion as scientific and pronounce it dependable. On the other hand, if the data are obviously subjective, the conclusion is called unscientific and is considered lacking in dependability. This tendency is not justified. The faults of data were pointed out in Chapter V and their significance in the study of certain types of problems was emphasized in Chapters IX, X, and XI. Objectivity of data does not guarantee dependability of conclusions. When the data are accurate, valid, and adequate with respect to the requirements of the problem, a conclusion derived from them is dependable. When the data are not entirely satisfactory, a conclusion may be shown to be dependable in spite of their limitations. Frequently, however, conclusions from research employing objective data must be considered lacking in dependability.

A conclusion reached by the methods of science is called dependable when it appears highly probable that a repetition of the investigation would lead to essentially the same conclusion. When applied to conclusions reached by means of the methods of philosophy, the meaning of dependability requires some adaptation. The criteria of dependability are implied in the questions: "Was the problem adequately defined?" "Were the basic assumptions recognized and understood?" "Were any pertinent data overlooked?" "Was the thinking biased or uncritical?" "Were the probable consequences of the conclusion adequately considered?" The philosopher is handicapped in testing his own thinking and even a competent critic may fail to detect weaknesses. It is difficult to foresee the consequences of a conclusion. In some cases they are revealed only as time passes. Hence, the testing of philosophical thinking frequently extends into the future. Sometimes the completion of the test

is deferred a generation or more. For example, the determination of the soundness of the conclusion that ability grouping should prevail in our schools probably cannot be completed until the resulting effect upon society one or more generations hence becomes apparent.

Since dependability of conclusions in philosophy does not mean the same as dependability of findings in science, a comparison can have only a limited significance. It seems reasonable, however, to say that the conclusions reached by a competent and critical person, relative to a problem of purposes, may be highly dependable and deserving of as much confidence as we give to reports of scientific findings in the field of education.¹

Objective techniques and the problem of curriculum construction.² Since curriculum construction furnishes a large number of the questions that ask what should be, the application of objective techniques in this field will be treated more explicitly. Reference is frequently made to "scientific curriculum construction" and examination of current curriculum studies reveals many applications of objective techniques. A curriculum consists of the objectives or goals pupils are expected to attain

¹ For an extended consideration of the question of the dependability of conclusions in philosophical inquiry, see Lepley, Ray. "Dependability in Philosophy of Education," *Teachers College, Columbia University Contributions to Education*, No. 461. New York: Bureau of Publications, Teachers College, Columbia University, 1931. 96 pp. The final chapter in which the author states his conclusion is worthy of careful study by anyone interested in the question.

² Douglass has reported an illuminating analysis of curriculum research in secondary education during 1929. By means of correspondence with instructors in colleges and universities and the examination of publications for the year he located 74 curriculum researches at the high school level, 40 of which were masters' theses. Approximately half of the studies dealt explicitly with the question of what the curriculum should be. With reference to the quality of the research, he states that "To a large extent the contributions of research to the secondary school curriculum have as yet been trivial and have been based on six conceptions which themselves are not established on scientific foundations and which to many educators seem definitely unsupportable." Douglass, H. R. "Types and Fields of Curriculum Research in Secondary Education during 1929," *School Review*, 38: 656-62, November, 1930.

The interested reader will find a good general discussion in the following reference:

Judd, C. H. "The Place of Research in a Program of Curriculum Development," *Journal of Educational Research*, 17: 313-23, May, 1928.

and of the learning exercises and materials of instruction to be employed as a means for securing the learning activities whose outcomes will be the abilities specified as immediate or control objectives. Usually, however, curriculum construction has been interpreted to mean only the determination of objectives. Since learning exercises and materials of instruction are essentially means, the determination of objectives is the basic problem.

When attempting a determination of objectives, a distinction should be made between remote or ultimate objectives and immediate objectives. Ultimate objectives are to be conceived of in terms of the characteristics (conduct) of the social group we desire to build up and perpetuate. For example, if a curriculum is being constructed for training plumbers, the ultimate objectives would be expressed in terms of the characteristics of the ideal plumber. Judgment cannot be avoided in determining this ideal plumber. The determination may be given the appearance of science by employing techniques of activity analysis, but judgment is introduced in selecting the plumbers whose activities are analyzed.

The techniques of activity analysis enable research workers to collect information useful in the formulation of ultimate objectives. However, judgment may be required in adding objectives which the techniques of activity analysis fail to discover. For example, objectives which have to do with the contemplation of the beautiful and the good and those which may be postulated as more important in the future than in the lives of adults of the present generation will not be revealed by activity analysis. Another limitation of activity analysis, especially when applied to adults, is that it neglects the activities of pupils which are necessary for an effective organization of the curriculum. For example, adults may be found to use a certain limited number of arithmetical operations, but children may need to use many more arithmetical operations during the course of their formal education in order to acquire and retain those that they will use as adults. The contention may be supported that activity analysis is a useful technique for col-

lecting needed factual data, but it is a technique which must be used with full recognition of its limitations and with the recognition that it cannot be expected to accomplish the determination of ultimate objectives without supplementation.¹

The techniques of activity analysis. Six techniques of activity analysis have been employed.

1. "Introspection," in which a participant in the activity lists all of the subsidiary activities or duties of which he can think.
2. "Working on the job," which is a modified form of introspection.
3. "Interviewing," in which a trained interviewer asks a participant in the major activity to give a list of his duties.
4. "Questionnaire," which is essentially a type of interviewing.
5. Observing workers and noting the particular duties they perform.
6. Analyzing records of activities performed.

Introspection is an effective technique in activity-analysis when the analyst has had considerable experience with the activity. Working on the job is useful where the analyst needs to acquire experience with the activity. The questionnaire and interview techniques are useful in securing data with respect to activities from a representative group of individuals. With the exception of representativeness, the data secured by all four of these techniques are similar in character. All involve introspection, and hence are subjective. The individual reporting the activities in which he is engaging, or has engaged, is likely to report those activities which are more or less routine in character to the neglect of activities that are performed only occasionally. These occasional activities may be the ones which require the greatest ability. The individual reporting his activities may tend to record those which seem to him important and neglect to report activities of whose importance he is unaware. For example, an auto mechanic might list comprehensively the various types of repair jobs in which he has engaged and neglect to mention his activities in dealing with customers which should be included if the analysis of his vocation is to be complete. Furthermore, he may report his activ-

¹For an elaboration of this point see Bode, B. H. *Modern Educational Theories*. New York: The Macmillan Company, 1927, Chapter V.

ities without indicating how they are performed. For example, the activity reported as "replacing broken timing gear" may represent insufficient analysis and inadequate information with respect to other activities engaged in concomitantly. A more complete analysis might include: responding promptly to motorist's telephone call for aid; towing car to town with precautions to prevent accident; diagnosing correctly what is wrong with the car; removing hood and radiator with care to prevent marring; jacking up engine; removing fan belt, fan, and gear casing; removing broken gear with suitable tools; selecting the appropriate new part; adjusting the new gear so that the timing will be correct; and so on, in detail, until the list is concluded with mention of a courteous word of thanks to the motorist and the request to call again.

The technique of observing workers has the advantage that the analyst may give more concentrated attention to the analysis of the activity. The worker may not be capable of effective introspection with respect to his job while performing it efficiently. The observer is in a somewhat better position to note the relations between the activities performed and the reactions of people served by the activities. The observer may be aided in his analysis by the utilization of such apparatus as a stopwatch and motion picture camera. The motion picture camera is particularly effective where a permanent record is desired which may be subjected to a detailed analysis of the motions used in performing the activity. The data obtained as a result of observation may be relatively objective in character.

The analysis of records of activities is a useful technique for obtaining information with respect to arithmetical operations performed by workers in various fields, words employed in written correspondence, and the like. It is also an advantageous technique in the analysis of activities other than vocational or professional. For example, analysis of the records of book withdrawals in public libraries yields information relative to the types of reading done by adults and by children. News-

papers and periodicals may be analyzed for the occurrence of certain types of items. Records of problem solutions in chemistry may be analyzed for the arithmetical operations needed in their solution.

The method of consensus of opinion.¹ The method of activity analysis is most useful when objectives are being determined for a well-defined occupation or activity. Its usefulness is limited when the problem is to determine the objectives of such school subjects as reading, geography, history, algebra, or physics. Expressions of opinions may be secured by means of a formal questionnaire, but some of the limitations of this procedure can be overcome by seeking out opinions that have already been expressed by competent persons. The study by Hockett described briefly in Chapter I is a good illustration of the latter method. It seems reasonable to assume that the problems and issues identified by Hockett do represent with a high degree of accuracy the problems and issues confronting the present generation of adults, and that many of these problems and issues will be applicable to the next generation.

The method of consensus of opinion as employed by Hockett is useful for determining ultimate objectives in several subject-matter fields. In many cases it enables the curriculum builder to set up objectives more likely to be valid in the preparation of children for the activities of adult life, than objectives determined merely through analysis of the activities of the present generation. It should be noted, however, that a consensus of opinion is likely to be weighted by tradition. Following the report of the Committee of Ten, curriculum construction by committees has been a common procedure. The high degree of agreement frequently reflected by the report is an indication of the influence of tradition upon their thinking.²

¹ For a discussion of an assumption basic to a consensus of opinion, see page 255.

² For an elaboration of this point see

Kelley, T. L. *Scientific Method*. New York: The Macmillan Company, 1932, pp. 152-64.

For a general criticism of consensus of opinion in curriculum construction see Bode, B. H. *Modern Educational Theories*. New York: The Macmillan Company, 1927. Chapter IV.

Techniques useful in the later stages of curriculum construction. After the ultimate objectives have been determined, other problems require attention: the determination of immediate objectives, the abilities necessary for performing the activities recognized as ultimate objectives; formulation of learning exercises that will be instrumental in stimulating and directing children in their acquisition of the abilities specified as immediate objectives;¹ the selection of suitable materials of instruction; and the determination of the placement of learning exercises and materials of instruction in an effective sequence in the organized curriculum. These problems call for the determination of relative merits of comparable means and hence may be attacked experimentally. The total program of required experimentation would be an extensive one and in the case of the determination of immediate objectives, several years would be required to complete an experiment. Our schools are in operation and when the curriculum revision is decided upon it seldom seems feasible to plan for a group of experimental studies extending over a period of years. Consequently, the curriculum for the next year is constructed by other means. Experimental studies, however, are steadily contributing information which makes the construction of the curriculum less subjective.

Experimentation in the psychological laboratory and under school conditions has contributed information relative to the nature of pupil abilities in several subject-matter fields.² For example, a number of studies have dealt with the specificity of the calculation skills of arithmetic, and, although the findings are not entirely consistent, they may be labeled an important contribution in the field of curriculum construction. The laboratory experimentation on the nature of silent reading ability conducted at the University of Chicago has had a very significant influence on the reading curriculum, particularly on the learning

¹ This problem includes the optimum adjustment of learning exercises to individual differences.

² Research involving the application of correlation analysis described in Chapter XI is also contributing to this end.

exercises and materials of instruction. There have also been numerous experimental studies of the relative merits of types of learning exercises and their organization. Among those studied are learning exercises relating to phonics, requests to learn rules in spelling, exercises in formal grammar, practice exercises in arithmetic, written assignments, the Dalton Plan and other contract organizations of learning exercises, projects, and lecture-demonstrations. A small number of experimental studies have dealt with the placement of learning exercises and materials of instruction.

Although the experimentation relating to the curriculum has been fragmentary and the number of dependable findings is not large, the possibility of such research is apparent and it seems reasonable to expect that eventually the curriculum-maker will have available sufficient experimental information to make the later stages of curriculum construction highly objective.¹ It should be noted, however, that the determination of the ultimate (conduct) objectives in which philosophical methods must be employed is required as a basis for such experimental studies.

Until experimental findings are much more adequate, the determination of immediate (control) objectives, the selection of learning exercises, and their organization and placement must be accomplished mainly by other means. Surveys of present practices and of opinion, including the analysis of textbooks and courses of study, will be helpful, but the limitations of the findings from such studies should be recognized. A consensus of opinion, as well as the average of present practices, is weighted by tradition. The findings of studies of pupil interests and of pupil achievement are influenced by present practices. However, the curriculum-maker who uses survey findings intelligently will probably formulate a better curriculum than he can make by disregarding such information.

¹ The range of scientific studies contributing to curriculum construction is wide. For an indication of their scope, see Thorndike, E. L. "Curriculum Research," *School and Society*, 28: 569-76, November 10, 1928.

BIBLIOGRAPHY

The items in this bibliography have been selected to provide references to discussions of curriculum construction and to illustrations of research techniques that may be employed in such work. The reader who is interested in the curriculum as a general topic will find it helpful to consult Dech, A. O. "A Guide to the Literature of the Curriculum," *Teachers College Record*, 35: 407-14, February, 1934.

AYRES, L. P. "The Spelling Vocabularies of Personal and Business Letters," *Pamphlet* No. E. 126. New York: Russell Sage Foundation, 1913. (Out of print, but reviewed in Charters, W. W. *Curriculum Construction*. New York: The Macmillan Company, 1923, p. 171.)

This is a significant pioneer study.

BAGLEY, W. C. *Education, Crime, and Social Progress*. New York: The Macmillan Company, 1931. 150 pp.

This monograph is worthy of critical study by the student interested in curriculum problems. The following chapter titles are indicative of its contents: "Some Handicaps of Character Education in the United States," "Discipline and Dogma," "Shibboleths and Slogans in Educational Reform," "Playing at the Work of Education," "Through Discipline to Freedom," "Emergent Idealism," and "Education for Adaptability."

BAGLEY, W. C., and RUGG, H. O. "The Content of American History as Taught in the Seventh and Eighth Grades," *University of Illinois Bulletin*, Vol. 13, No. 51, *School of Education Bulletin*, No. 16. Urbana: University of Illinois, 1916. 59 pp. (Out of print.)

Twenty-three American history texts ranging over the period 1865 to 1915 and intended for use in the seventh and eighth grades were selected in a "random" fashion. Analysis resulted in topics and names common to all of the books, those common to at least three-fourths of the books, those common to at least half of them, and finally the amount of space devoted to each topic. This is a significant pioneer study.

BAMESBERGER, V. C. "An Appraisal of a Social Studies Course, in Terms of Its Effect upon the Achievement, Activities, and Interests of Pupils," *Teachers College, Columbia University Contributions to Education*, No. 328. New York: Bureau of Publications, Teachers College, Columbia University, 1928. 91 pp.

While this study classifies as an experiment, inclusion in the present bibliography is appropriate since it illustrates a logical "final" step in curriculum research.

BOBBITT, FRANKLIN. *The Curriculum*. Boston: Houghton Mifflin Company, 1918. 295 pp.

The first systematic treatise in the field of curriculum construction.

BOBBITT, FRANKLIN. *How to Make a Curriculum*. Boston: Houghton Mifflin Company, 1924. 292 pp. And Bobbitt, Franklin, et al. "Curriculum-Making in Los Angeles," *Supplementary Educational Monographs*, No. 20. Chicago: University of Chicago Press, 1922. 106 pp. (Out of print.)

Over a period of twelve years, several hundred objectives were collected by Bobbitt and some fifteen hundred members of graduate classes in "The Curriculum." The tentative list was submitted to "citizens, school officials, and teachers of Los Angeles." The critical examination made by some twelve hundred high school teachers formed the chief basis of revision. The final list represents a consensus of opinion.

BOBBITT, FRANKLIN, et al. "Curriculum Investigations," *Supplementary Educational Monographs*, No. 31. Chicago: University of Chicago Press, 1926, pp. 7-22.

In this study, the *Reader's Guide to Periodical Literature* for the three-year period of 1919-1921 was analyzed for the purpose of discovering the "major activities of man's life" and the "subordinate fields into which the major fields naturally divide themselves." It is concluded that the data given in the tables "go a long way toward showing the things which function in human life *today*."

In the later chapters of the monograph are given analytical studies whose nature is indicated by the titles: "Major Fields of Human Concern: The Evidence from the Literary Digest" (P. L. Palmer), "Duties and Traits of a Good Citizen" (J. A. Nietz), "Civic and Social Shortcomings as Curriculum Indices" (I. H. Dulebohn), "Social Problems of the Labor Group" (G. K. Bixler), "Quality of Conduct" (Franklin Bobbitt, et al.), "Approved Social Behavior" (C. H. Lorenzen), "Shortcomings of the Written English of Adults" (Sarah Bobbitt), "The Mathematics Used in Popular Science" (R. C. Scarf), "Play Activities of Persons of Different Ages" (H. C. Lehman), and "The Placement of Poems in the Grades" (C. A. Dyer).

BODE, B. H. *Modern Educational Theories*. New York: The Macmillan Company, 1927. 351 pp.

The author shows the impossibility of determining *what should be* merely by collecting factual descriptions of *what is*, or by collecting opinions. A biting criticism of objective methods in curriculum construction by a leading educational theorist.

BOWDEN, A. O. "Consumers' Uses of Arithmetic, an Investigation to Determine the Actual Uses of Arithmetic in Adult Social Life, Exclusive of Vocational Uses," *Teachers College, Columbia University Con-*

tributions to Education, No. 340. New York: Bureau of Publications, Teachers College, Columbia University, 1929. 69 pp.

The problem of this investigation is indicated in the title. See criticism by W. J. Osburn. "Two Recent Books on Arithmetic," *Educational Research Bulletin* (Ohio State University), 9: 66-73, February 5, 1930.

BRIGGS, T. H. *Curriculum Problems*. New York: The Macmillan Company, 1926. 138 pp.

Twenty-seven fundamental questions of the curriculum are listed and their implications discussed. The latter part of the book is devoted to discussions of emotionalized attitudes and mores ("the manner of action generally accepted in a social group") with emphasis on their significance to the problems of curriculum construction.

CHARTERS, W. W. *Curriculum Construction*. New York: The Macmillan Company, 1923. 352 pp.

This is an important treatise on the techniques of curriculum construction.

CHARTERS, W. W., and WAPLES, DOUGLAS. *The Commonwealth Teacher Training Study*. Chicago: University of Chicago Press, 1929. 666 pp.

A master list of teachers' activities was compiled in this study from lists prepared by other investigators, from lists submitted by professional educators, and from blanks filled out by several thousand teachers. Professors of education and their graduate students were asked to list activities which teachers *ought* to perform. The master list includes under seven major heads, twelve thousand activities. This research constitutes a basic contribution in the field of teacher training curricula.

CHARTERS, W. W., and WHITLEY, I. B. "Summary of Report on Analysis of Secretarial Duties and Traits," *Service Bulletin*, No. 1. New York: National Junior Personnel Service, Inc., 1924. 62 pp.

Trained workers, using a series of carefully prepared questions, interviewed 125 secretaries. Eight hundred seventy-one duties were discovered. The interview technique was supplemented by a questionnaire in which the 871 duties were submitted to secretaries and stenographers for checking. The relative frequency of the 871 duties was ascertained from a tabulation of 715 checked duty lists. Relative frequency of duties was determined not only for the total group of 715 secretaries and stenographers, but also for each of fifteen groups into which the 715 were divided according to the profession or business of the employers. See the reference to the research of Tyrrell in this connection.

COCKING, W. D. "Administrative Procedures in Curriculum Making for Public Schools," *Teachers College, Columbia University Contributions to*

Education, No. 329. New York: Bureau of Publications, Teachers College, Columbia University, 1929. 120 pp.

This monograph contains a comprehensive history of curriculum making in which the influences of European schools, colleges, national committees, public opinion, state legislatures, surveys, philosophers, and scientific method are indicated. Chapters are devoted to the underlying principles of curriculum making, procedures for instituting a program of curriculum construction, part to be played by boards of education and other groups, committee organization, and ways of evaluating and appraising curriculum work.

COUNTS, G. S. "The Senior High School Curriculum," *Supplementary Educational Monographs*, No. 29. Chicago: University of Chicago Press, 1926. 160 pp.

Data were collected with respect to the curricula of high schools in fifteen cities of the United States.

DEWEY, JOHN. *The School and Society*. Chicago: University of Chicago Press, 1900, 1915. 164 pp.

This is but one of several books of John Dewey which bear on the problems of the curriculum. Other titles are: "The Child and the Curriculum," "Interest and Effort in Education," and "Democracy and Education." The student will find it profitable to study also "Essays in Experimental Logic," "Experience and Nature," "The Quest for Certainty," and "Sources of a Science of Education."

EASON, J. L. "Diagnostic Study of Technical Incorrectness in the Writing of Graduates of Tennessee County High Schools," *Contributions to Education*, No. 64. Nashville: George Peabody College for Teachers, 1929. 89 pp.

A total of 638 freshmen, entering the University of Tennessee from a representative sample of Tennessee high schools, were requested to write 300 word themes on three topics. The papers were analyzed for technical and mechanical errors by several teachers of freshman English.

FINLEY, C. W., and CALDWELL, O. W. *Biology in the Public Press*. New York: Lincoln School of Teachers College, Columbia University, 1923. 151 pp.

Four hundred and ninety-two daily papers were analyzed with reference to biological articles.

FULLER, L. R. "Manual Arts Based on Home Repair," *Journal of Educational Research*, 3: 173-79, March, 1921.

This investigator undertook "to determine what problems and processes would be involved in a manual arts course, based upon work which is done

or may be done around the home by a handy man with common carpenter's or painter's tools." Data were collected by means of observation, interview, and questionnaire.

HARAP, HENRY. *The Education of the Consumer*. New York: The Macmillan Company, 1924. 360 pp.

In this study, Harap sought to discover the objectives relating to the consumption of food, shelter, fuel, and clothing. Data were collected from the Census Reports, United States Bureau of Labor statistics, and reports of independent studies.

HARAP, HENRY. "A Critique of Public School Courses of Study, 1928-29," *Journal of Educational Research*, 21: 109-19, February, 1930.

An analytical survey of 242 courses of study which in the opinion of the investigator represented approximately one-half of all produced in the country during the years 1928 and 1929.

HARAP, HENRY. *The Technique of Curriculum Making*. New York: The Macmillan Company, 1928. 315 pp.

An important text on the techniques of curriculum construction. A good bibliography is given.

HARAP, HENRY, and PERSING, E. C. "The Present Objectives in General Science," *Science Education*, 14: 477-97, March, 1930.

Thirteen leaders in the teaching of science evaluated and suggested the sources analyzed. These sources included 5 curriculum investigations, 11 courses of study, and 5 texts.

HOCKETT, J. A. "A Determination of the Major Social Problems of American Life," *Teachers College, Columbia University Contributions to Education*, No. 281. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 101 pp.

See the description to this study given in Chapter I, pages 6-7. The following study reported by Horn is a pioneer investigation of the same type: Horn, Ernest. "Possible Defects in the Present Content of American History as Taught in the Schools," *Sixteenth Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1917, pp. 156-72.

HOPKINS, L. T. *Curriculum Principles and Practices*. Chicago: Benjamin H. Sanborn and Company, 1929. 617 pp.

This is a very comprehensive treatise on curriculum construction dealing with both philosophy and techniques.

KILPATRICK, W. H. *Education for a Changing Civilization*. New York: The Macmillan Company, 1926. 143 pp.

Advances the thesis that civilization is changing too rapidly to plan a curriculum in advance.

MAHAN, T. J. "An Analysis of the Characteristics of Citizenship," *Teachers College, Columbia University Contributions to Education*, No. 315. New York: Bureau of Publications, Teachers College, Columbia University, 1928. 44 pp.

This investigator collected opinions of 350 high school pupils with respect to the characteristics of citizenship. He also collected opinions and experiences with respect to citizenship by means of interview and questionnaire from 640 "representative" citizens of the United States. In addition, he analyzed five commonly used civic texts to determine the extent to which they treat the specific duties, difficulties, and qualities named by the representative citizens.

MERIAM, J. L. *Child Life and the Curriculum*. Yonkers-on-Hudson, New York: The World Book Company, 1920. 538 pp.

The author criticizes traditional curricula and discusses curriculum construction in which materials and activities are developed directly from the pupil's out of school interests. Meriam's theory was applied by Collings in his experimental rural school. Collings, Ellsworth. *An Experiment with a Project Curriculum*. New York: The Macmillan Company, 1923. 346 pp.

MONROE, W. S., and CLARK, J. A. "The Teacher's Responsibility for Devising Learning Exercises in Arithmetic," *University of Illinois Bulletin*, Vol. 23, No. 41, *Bureau of Educational Research Bulletin*, No. 31. Urbana: University of Illinois, 1926. 92 pp.

The problem of this research was "to determine the nature and extent of the learning exercises provided by texts in arithmetic." The analysis resulted in 333 problem types. These were used as a basis for analyzing the second and third books of ten three-book series of arithmetics.

MONROE, W. S., and HERRIOTT, M. E. "Reconstruction of the Secondary School Curriculum: Its Meaning and Trends," *University of Illinois Bulletin*, Vol. 25, No. 42. *Bureau of Educational Research Bulletin*, No. 41. Urbana: University of Illinois, 1928. 120 pp.

This is a comprehensive study of the trends in the reconstruction of the secondary-school curriculum during the period 1893-1928.

MONROE, W. S., HINDMAN, D. A., and LUNDIN, R. S. "Two Illustrations of Curriculum Construction," *University of Illinois Bulletin*, Vol. 25,

No. 26, *Bureau of Educational Research Bulletin*, No. 39. Urbana: University of Illinois, 1928. 53 pp.

The illustrations of curriculum construction described in this monograph involve little or no use of objective data. The procedure employed may be described as "systematic and critical judgment." In both cases, the first major step was to formulate an analytical description of the ultimate or conduct objectives for which the proposed curriculum was considered to contribute equipment. From these conduct objectives, the immediate or control objectives were derived. The last two steps include the predicting of the learning activities necessary for acquiring the specified controls of conduct, and the learning exercises that will serve as efficient bases for these activities.

MUTHERSBAUGH, G. C. "Objectives of a Proposed Course of Study in Physics for Senior High Schools," *School Science and Mathematics*, 29: 943-54, December, 1929.

Four texts in high school physics and four courses of study were selected for analysis on the basis of the judgments of 16 leaders in the field of physics teaching. These sources were examined for objectives, an objective being defined as "a specific goal expressed in terms of useful life situations." The analysis resulted in 1018 slips each containing a stated objective. Classification and elimination of duplicates reduced the number to 275. A number were then eliminated on the basis of inapplicability to useful life situations after a rating on a ten-point basis with respect to usefulness, frequency of occurrence, and interest. The 221 objectives finally presented in the report are classified under 43 "units." The author has added a list of 20 supplementary objectives. The study exemplifies the inadequacy of "objective" data in curriculum research. Judgment was introduced in the elimination and supplementation.

PEIK, W. E. "The Analysis and Evaluation of College and University Courses in Education," *Journal of Educational Research*, 18: 345-55, December, 1928.

Syllabi or complete sets of lesson units submitted by thirteen instructors of fifteen educational courses including eight in special methods were analyzed into 814 topics of instruction. These topics were evaluated by a group of alumni with respect to helpfulness in classroom teaching and educational thinking. The alumni were also asked to indicate which topics they remembered having been taught in the prescribed courses, which should be omitted from such courses, and which were treated inadequately.

PETERS, C. C. *Objectives and Procedures in Civic Education*. New York: Longmans, Green and Company, 1930. 302 pp.

Chapter IV contains a "Blue Print of an Optimum Citizen" made up of short statements of the objectives of education for citizenship as derived from over a thousand separate studies made by the author and his students. Included are suggestions of possible means and occasions for training to meet these objectives. The student should read in this connection Kilpatrick's criticism and Peter's defense of his philosophy and his method: Kilpatrick, W. H. "Hidden Philosophies," *Journal of Educational Sociology*, 4: 59-68, September, 1930. Peters, C. C. "Revealed Philosophies—A Reply to Professor Kilpatrick," *Journal of Educational Sociology*, 4: 260-71, January, 1931. Peters holds that the conflict between the opposing schools of curriculum theorists and builders is a conflict of philosophies. The pragmatists, Dewey, Kilpatrick, and Bode stress the world in the making while Peters, whose philosophy is that of absolute idealism, does not feel that the world is too chaotic to prevent systematic planning of curricula.

REAGAN, G. W. "The Mathematics Involved in Solving High School Physics Problems," *School Science and Mathematics*, 25: 292-99, March, 1925.

A total of 241 problems in Millikan and Gale's *A First Course in Physics* were solved and analyzed. The results are reported under the headings of arithmetic, algebra, and geometry.

RUGG, H. O., et al. "The Foundations and Technique of Curriculum-Making," *The Twenty-Sixth Yearbook of the National Society for the Study of Education*. Bloomington, Illinois: Public School Publishing Company, 1926. Part I, 475 pp.; Part II, 210 pp.

Part I, "Curriculum-Making Past and Present deals with the historical development of the curriculum and present practices in curriculum construction. Sections are devoted to examples of curriculum construction in progressive public school systems and in private laboratory schools. The last section of this volume contains a review and critique of curriculum-making for the vocations, curriculum reconstruction on the college level, curriculum-making by state legislatures, an appraisal of current methods of curriculum-making, and an extensive annotated bibliography.

Part II, "The Foundations of Curriculum-Making" presents a compilation of principles with respect to the curriculum signed by the members of the committee. This compilation is followed by supplementary statements of the committee members indicating their individual points of view with respect to the principles previously expressed. The volume concludes with quotations on the curriculum from the writings of John Dewey and a number of quotations from the Herbartians and their critics.

STRATEMEYER, F. B., and BRUNER, H. B. "Rating Elementary School Courses of Study," *Studies of the Bureau of Curriculum Research of*

Teachers College, Columbia University, Bulletin No. 1. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 193 pp.

In this monograph are described the techniques used in rating eight hundred to a thousand courses of study in each of the subject-matter fields of the elementary school. One hundred twenty-one judges participated using criteria, formulated as rating scales, and derived from examination of a large number of courses of study. These criteria constitute one of the chief contributions of the study. A list of courses of study are given which most nearly conform to the "best points of the criteria."

STRONG, E. K. "Job Analysis of the Manager of Industry," *School and Society*, 13: 456-62, April 16, 1921.

The author discusses research conducted at Carnegie Institute of Technology in which job analyses were made in three fields of executive work—building construction, commercial printing, and metal working industries.

TYRRELL, DORIS. "An Activity Analysis of Secretarial Duties as a Basis for an Office Practice Course," *Journal of Experimental Education*, 1: 323-40, June, 1933.

In this study an evaluation is reported of 406 secretarial duties given in the list compiled in previous research by Charters and Whitley. In addition to the frequency ratings of Charters and Whitley, here reported in terms of decile ranks, the author reports evaluations with respect to the importance, difficulty, and desirability for pre-service training of each of the duties. This study represents a stage in the process of curriculum construction midway between the work of activity analysis and the formulation of a course of study.

WILSON, G. M. "A Survey of the Social and Business Use of Arithmetic," *Sixteenth Yearbook of the National Society for the Study of Education*. Part I. Bloomington, Illinois: Public School Publishing Company, 1917, pp. 128-29. *See also:*

WILSON, G. M. *What Arithmetic Shall We Teach?* Boston: Houghton Mifflin Company, 1926, pp. 7-9, 30-51, 58-63.

The purpose of this research was to determine the arithmetic "actually needed by social and business usage." Sixth-, seventh-, and eighth-grade pupils were asked to collect "every problem solved by either the father or the mother . . . through a period of two weeks." This is an important pioneer study.

WRAY, ROBERT P. "The Relative Importance of Items of Chemical Information for General Education," *Journal of Experimental Education*, 1: 341-89, June, 1933.

A list of 1550 items of chemical information, secured as the result of an analysis of four commonly used high school chemistry texts, was evaluated in this study. In obtaining the evaluations, fifteen questionnaires of the check-list type were sent to various groups of individuals. For example, the first questionnaire was returned by 176 persons including teachers, engineers, students, laborers, housekeepers, medical men, secretaries, and business men.

CHAPTER XIII

EVALUATING AND SYNTHESIZING EDUCATIONAL RESEARCH

The need for evaluating reported conclusions. The preceding chapters have dealt primarily with the production of educational research. Consumers of research, however, are much more numerous than those engaged in production. Hence, it is appropriate that we consider also the problems arising in the use of reported research. An important question relates to the acceptance of the stated conclusions at their face value. Is the reader justified in accepting conclusions as reported, or should he seek to determine for himself their dependability?

When considering this question one should bear in mind that educational research is a recent development. The first text on statistical methods applied to educational problems, Thorndike's *Theory of Mental and Social Measurement*, was not published until 1904, and the development of instruments for measuring mental traits and abilities had scarcely begun by 1910. Although much has been accomplished in the development of research techniques, especially since 1918, there are many unsolved problems in the field of educational measurements, and none of the many texts on statistical methods provides a satisfactory treatment of statistical techniques. In view of these conditions it is to be expected that some of the reported conclusions are not dependable. The popularization of educational research and the attainment of quantity production have operated to increase greatly the amount of amateurish activity. The emphasis upon the merits of objective data has tended to cause investigators to neglect the limitations of the data being used and the employment of statistical techniques that were only vaguely under-

stood ¹ has encouraged a mechanized interpretation, especially in the case of differences and of coefficients of correlation.

A number of critical writers have commented upon the quality of educational research,² pointing out that reported conclusions are not infrequently lacking in dependability. Sometimes the research has been done carelessly without a well defined problem. In other cases the data are inaccurate or incomplete, and hence, even though the author has endeavored to apply appropriate statistical techniques his findings are subject to qualifications and limitations which have not been adequately noted in his report. Occasionally the statistical treatment is not appropriate and in a surprising number of cases there has been failure to interpret statistics correctly and with precision. Hence, a reader should be critical and not accept, without evaluation, reported conclusions.

The need for summarizing reported research. The quantity production that has prevailed in educational research since soon after 1920 has resulted in a large number of investigations relating to many topics and problems. Today there is scarcely a topic for which it is not possible to compile an extensive bibliography of references purporting to be reports of research. In many cases a bibliography relating to a fairly narrow topic will include more than a hundred items. Hence a person interested in the results of research relating to a particular topic or problem frequently faces a task that requires many hours of labor. When a comprehensive bibliography is examined it is not infrequently found that several references are to unpublished studies. Others are to bulletins or periodicals not easily accessible, especially to the practical schoolman. Even when the references listed are accessible the task of reading a large number of reports requires time. The practical schoolman seldom has the time or, if he does, the task does not appear to be the most worthy one com-

¹ Lehman, H. C., and Witty, P. A. "Statistics Show—," *Journal of Educational Psychology*, 19: 175-84, March, 1928.

Although reference is not made to particular studies, this article is a severe indictment of the statistical aspects of educational research.

² For illustrations of critical statements, see Chapter XIV, page 467.

manding his attention. Hence, he selects a few of the most attractive and most accessible references and accepts the findings of these studies as representing the results of research relating to the problem in which he is interested. Graduate students and other persons, except the more systematic and conscientious research workers, frequently follow about the same plan. The need for summaries of the results of research is therefore apparent.

In addition to facilitating the use of educational research, summaries are helpful to the research worker. Critical summaries make evident the phases of problems that have been inadequately studied thereby assisting an investigator in defining his own problem. If the summary includes comments relative to the advantages and disadvantages of the techniques employed, he will be helped in deciding upon the details of his procedure. In case a critical summary is not available the first step in planning a research should be a review of the studies relating to the proposed problem. Not infrequently the number of such studies is so great that the task of evaluating them and synthesizing the findings consumes so much time that the investigator's enthusiasm is greatly diminished. When his time is limited, as in the case of graduate students, he finds that when the summary is completed he has little or no time for his own inquiry.

Published summaries. The need for summarizing the results of educational research is being recognized. In 1925, W. S. Gray published a summary of investigations in the field of reading which has been supplemented annually. In 1925, Buswell and Judd published a summary of investigations in arithmetic which also has been supplemented annually by Buswell. There are a number of other extensive summaries, notably that of Curtis, *A Digest of Investigations in the Teaching of Science*, which appeared in 1926 and was supplemented in 1931; and Lyman, *Summary of Investigations Relating to Grammar, Language, and Composition* which appeared in 1929. Many reports of research include a summary of previous summaries relating to the prob-

lem. In February, 1931, the American Educational Research Association began the publication of the *Review of Educational Research* which will present once every three years a summary of the research relating to the following general topics.¹

1. Methods and Techniques of Educational Research.
2. The Curriculum.
3. Teacher Personnel.
4. School Organization.
5. Psychology and Methods in the High School and College.
6. Special Methods and Psychology of the Elementary School Subjects.
7. Finance and Business Administration.
8. Psychological Tests.
9. Buildings, Grounds, Equipment, and Supplies.
10. Educational Tests.
11. Mental and Physical Development and Individual Differences.
12. Student Accounting, Personnel, and Guidance.
13. General Methods and Supervision.
14. History of Education and Comparative Education.
15. Legal Basis of Education.

The quality of published summaries. Unfortunately few of the summaries being published are critical. In most cases the reported conclusions of the studies listed are presented without evaluation. For example, in a recent summary of which the bibliography totaled 885 references, evaluative statements were made in the case of only 52 of the studies, and only 5 of the evaluative statements indicated that the reported conclusion was lacking in dependability or was subject to qualification. In view of the difficulties that are encountered in educational research and the inadequacies of our techniques, it is inconceivable that in such a comprehensive list of references there should not be a much larger number of researches whose reported findings deserve criticism. This summary is, perhaps, not typical, but an extended examination of published summaries indicates that only relatively few may be regarded as critical.² This is un-

¹ The list of topics for the first three-year cycle ending December, 1933, was slightly different. See page 462 for the list.

² For a systematic evaluation of two summaries, see Wilson, G. M. "Research: Suggested Standards for Summarizing and Reporting Applied to Two

fortunate because summarization without critical evaluation contributes to the perpetuation of error and triviality.

Evaluation and summary as a type of research. Although the summarization of pertinent studies should be a phase of any piece of research, it seems appropriate to propose evaluation and summary as a type of research, especially when the field considered is at all comprehensive. Such endeavors require the utilization of critical reflective thinking and a scientific attitude, the collection of the best data obtainable in a systematic manner, and the use of these data with full recognition of their limitations. The total procedure of preparing a critical summary conforms with the definition of educational research given in Chapter I. Although critical summaries, however comprehensive, can seldom be termed original contributions in the sense of providing new knowledge, when well done they contribute to the organization of the science of education. The evaluation, interpretation, and synthesis of the findings of original studies is an essential phase of the development of a science.

Phases of the preparation of a critical summary. The first step in the preparation of a summary is the compilation of a bibliography of researches relating to the problem or topic. The second step is to examine critically each of the researches to determine the dependability of the reported conclusions. If a conclusion as stated is not judged to be dependable the reviewer may be able to formulate one that appears to be justified by the data collected, or at least to supplement the stated conclusion with suitable limitations and qualifications. The final step is to organize and synthesize the dependable conclusions into a summary account of what research has revealed relative to the phases of the problem or topic.

Compiling a bibliography.¹ Published bibliographies are available for a large number of fields and topics and frequently a

Recent Summaries of Studies in Arithmetic," *Journal of Educational Research*, 28: 187-94, November, 1934.

¹ For a comprehensive treatment see Alexander, Carter. *How to Locate Educational Information and Data*. New York: Bureau of Publications, Teachers College, Columbia University, 1935. 272 pp.

worker will find several lists of references from which he may compile a bibliography for the topic or problem in which he is interested. The bibliographies of the *Review of Educational Research* include practically all of the better studies for the period covered and the selected lists of references in the *Elementary School Journal* and the *School Review* are helpful.¹ Other lists of references may be located by consulting a bibliography of bibliographies.² The *Education Index* initiated January, 1929, indexes by topic and author educational periodicals in English, a few of the more important educational periodicals in foreign languages, and current books, pamphlets, and documents. The *Psychological Index* published since 1894 is helpful in compiling a bibliography in the field of psychology.³ The publications of United States Office of Education include many bibliographies.⁴ Titles of doctors' and masters' theses in education for the period January, 1917, to October, 1927, are listed in compilations prepared by the Bureau of Educational Research at the University of Illinois.⁵ Since 1927, graduate

¹ The publication of these lists of selected references was begun in 1933. At the end of the year, the twenty bibliographies are assembled and published in a monograph with a title of "Selected References in Education."

² Monroe, W. S., and Asher, Ollie. "A Bibliography of Bibliographies," *University of Illinois Bulletin*, Vol. 24, No. 44, *Bureau of Educational Research Bulletin*, No. 36. Urbana: University of Illinois, 1927. 60 pp.

Monroe, W. S., Hamilton, T. T., and Smith, V. T. "Locating Educational Information in Published Sources," *University of Illinois Bulletin*, Vol. 27, No. 45, *Bureau of Educational Research Bulletin*, No. 50. Urbana, Illinois: University of Illinois, 1930, pp. 58-142. This bulletin includes the more important bibliographies listed in the earlier publication.

Monroe, W. S., and Shores, Louis. *Bibliographies and Summaries in Education*. New York: The H. W. Wilson Company, 1936. 465 pp. Over 4000 references are indexed.

³ For information in regard to other bibliographical aids, see Witmer, E. M., and Miller, M. C. "Guides to Educational Literature in Periodicals," *Teachers College Record*, 33: 719-30, May, 1932.

⁴ For information in regard to these publications, see Witmer, E. M., and Miller, M. C. "U. S. Office of Education Serial Publications," *Teachers College Record*, 34: 302-11, January, 1933.

⁵ The reference to the most recent compilation is Monroe, W. S. (Compiler). *Titles of Masters' and Doctors' Theses in Education Accepted by Colleges and Universities in the United States between October 15, 1925 and October 15, 1927*. Urbana, Illinois: College of Education, University of Illinois, 1928. 252 pp.

theses have been included in the *Bibliography of Research Studies in Education* published annually by the United States Office of Education. A number of institutions are publishing annotated lists of the theses accepted in partial fulfillment for graduate degrees in education. Other institutions publish merely lists of titles.¹

The graduate student in education should become familiar with the important journals, yearbooks, and monograph series² so that he will be able to locate likely sources of relevant references. Under the direction of Carter Alexander, Library Professor, Teachers College, Columbia University, a series of guides to the professional literature related to various phases of education have been prepared. These are being published in various periodicals so as to bring each article to the attention of those readers who are most likely to be interested in it.³ In compiling a bibliography it is frequently desirable to "thumb through" the pages of the journals that are known to contain reports of research in the field of the problem. If it does not seem worth while to look through the pages of the volumes, it is almost as effective to scan the indexes. "Thumbing the pages" is more effective since it overcomes the handicap of general or misleading titles.

The graduate student, or other research worker in education,

¹ See Derring, C. E. "Lists and Abstracts of Masters' Theses and Doctors' Dissertations in Education," *Teachers College Record*, 34: 490-502, March, 1933.

The worker who is interested in learning about research under way or completed but not published will find information in regard to sources in the following reference:

Witmer, E. M. "Educational Research: A Bibliography on Sources Useful in Determining Research Completed or under Way," *Teachers College Record*, 33: 335-40, January, 1932.

² For a comprehensive description of educational serials, see Monroe, Hamilton, and Smith, *op. cit.*, pp. 19-57.

³ The one for the field of secondary education was published in the *School Review*.

Manske, A. J., and Alexander, Carter. "Guide to the Literature on Secondary Education," *School Review*, 42: 368-81, May, 1934.

In addition to a selected list of bibliographies, information is given relative to sources of information for such topics as the following: periodicals, associations, book reviews, editorial comment, news notes, researches completed, under way, or needed, and statistics.

should be systematic in compiling a bibliography. He should use each of the aids suggested completely before going on to the next. That is to say, he should extract all the help the aid has to offer before looking elsewhere for references. It is desirable to keep a memorandum of the sources from which a bibliography has been compiled. The following is an example of such a record:

1. Education Index, January, 1929, to September, 1932.
2. Journal of Educational Research, January, 1920, to September, 1932.
3. Teachers College, Columbia University Contributions to Education, No. 300 to No. 545.

When such a record of sources is kept, it is a simple matter to bring the bibliography up to date. Furthermore, a record of the sources tends to stimulate the student to more systematic and scholarly effort.

In copying a reference, complete information should be recorded. The use of 3×5 or 4×6 cards is recommended. Larger cards, or even manuscript paper, are desirable when annotations are to be made.

Criteria for identifying educational research. It may appear unnecessary to raise the question of identifying educational research, but certain writers have pointed out that much of what is commonly called educational research does not deserve this label. For example, Rugg has said that "most of our so-called 'educational research' is not educational research at all."¹ The same position is taken in an editorial in the *School Review* for September, 1926, commenting on the "Bibliography of Secondary Education Research, 1920-25."²

Such criticisms imply criteria by means of which reports of

¹ Rugg, H. O. "Statistical Methods Applied to Educational Testing," *Twenty-First Yearbook of the National Society for the Study of Education*. Bloomington, Illinois: Public School Publishing Company, 1922, pp. 45-91.

² Windes, E. E., and Greenleaf, W. J. "Bibliography of Secondary Education Research, 1920-25," *U. S. Bureau of Education Bulletin*, 1926, No. 2. Washington, 1926. 95 pp.

educational research may be distinguished from writings that do not represent research, but no authoritative list has been formulated. Thoughtful consideration of the matter reveals the difficulty. No sharp line of demarcation can be defined. Although the extremes of both groups of writings stand out clearly, one group tends to merge into the other. The difficulty of identifying research is accentuated by reason of the fact that some investigators quit before their data have been adequately interpreted and that others are careless or inexperienced in the use of the techniques of research.

In identifying writings that should be labeled "research," it seems desirable to be rather highly selective. From the point of view of developing a science of education, studies that represent a mere collection of information or whose findings have little significance outside of the situation studied must be regarded as trivial. Such studies may be useful in the administration of a school or in other practical activities, but their contribution to a science of education is slight. Frequently there is no direct contribution. Hence, it seems desirable to label such investigations, "service studies" and to restrict the use of "research" to studies that have a distinct contributory value. It is from this point of view that the following criteria are suggested as a means of identifying research.

1. There should be a problem which functions as a guide in collecting data and in the subsequent phases of the work. Usually this problem is clearly defined by the investigator as a preliminary phase of his work.
2. An essential requirement is that the data collected afford some basis for generalization and that the interpretation of the data be continued until a tentative generalization is reached. As used here a generalization designates a statement of conditions, trends, or relationships which may be utilized as a basis of thinking or action in situations other than the particular one studied.
3. Another essential requirement is that in interpreting the data adequate recognition be given to their faults and to the limitations of the statistical procedures employed in handling them.

The first criterion affords a basis for eliminating discussions, expressions of opinions, propagandistic writings, and the like. There must be systematic collecting of data pertinent to a problem. Application of the second criterion will result in the rejection of studies in which the data collected are lacking in representativeness or for other reasons do not afford a basis for generalizing. Among the studies thus rejected will be those whose significance is only local, and those based upon too few cases,¹ and those in which no consideration has been given to the representativeness of the data. Other studies will be rejected because the author failed to continue his inquiry to the stage of generalization. Such studies may be useful to another investigator and if other requirements are satisfied may be appropriately labeled "incomplete research." Application of the third criterion will result in the rejection of studies in which the interpretation of the data has not been critical. Some of those thus rejected might be labeled "poor research," meaning thereby that the cause of rejection is the failure to use appropriate techniques or to use skillfully the techniques employed.

If a list of titles has been compiled without examining the writings, application of these criteria will usually result in the rejection of a large proportion of the items. Some of the references will be found to be essentially only discussions or descriptions of educational practice. A few are likely to be proposals for research. Still others will be found to be trivial.

Selecting research pertinent to one's problem. In addition to identifying the references that justify the label of research it is necessary to sort out those that are pertinent to one's problem or topic. A prerequisite for doing this is a precise definition of the problem or topic. A reference may be "interesting" but if it is not pertinent, failure to exclude it will add to the labor of evaluation. Hence, before a reference is read critically it should

¹ No definite number of cases can be prescribed as a prerequisite for generalization. If the area to which the generalizations apply is restricted, a small number of cases may be sufficient. For example, certain generalizations relative to highly gifted children might be justified from an intensive study of ten typical cases.

be examined to determine its pertinency to the problem or topic under consideration.

Evaluation of reported research. The purpose in evaluating a piece of educational research is to arrive at an estimate of the dependability of the author's conclusions. The best test of the dependability of a conclusion is to repeat the reported investigation duplicating the conditions and techniques as nearly as possible. This procedure, however, is seldom feasible. Hence, it is usually necessary for a reviewer to resort to a critical examination of the report and when possible to comparison with other studies of the same or closely related problems.

The evaluation of a reported study involves essentially the same considerations as the determination of dependability by the author to which attention has been directed in the preceding chapters, especially VIII, IX, X, and XI.¹ The reviewer, however, is handicapped by reason of the fact that he has only the information that the author has reported. He should seek the definition of the problem. If it is not given, he should endeavor to formulate a precise statement of it. Next, the general character of the data and the techniques employed in collecting them should be noted. The most important phase of the evaluation is to ascertain the probable faults of the data² and the compatibility of the stated conclusions with the data when their faults and limitations are considered.

There are few definite techniques that may be employed for identifying the faults in the data of a research. In general, one who attempts to evaluate a piece of research must rely upon his acquaintance with educational data and the techniques of educational research. An experienced person who is critically minded will usually be able to estimate the faults of the data. The reviewer should also examine the handling of the data and note any inappropriate procedures or incorrect interpretations of the findings. When possible, it is sometimes desirable to check calculations.

¹ For specific page references, see "dependability" in the index of this volume.

² For a general exposition of these faults and their significance, see Chapter V.

When the reviewer has two or more studies of the same problem or related problems, comparison of the reported conclusions may yield an indication of their dependability. Comparisons, however, must be made with caution and the fact that two reported conclusions are not in agreement does not necessarily mean that one of the conclusions is lacking in dependability. It is possible that the disagreement may be explained by differences in the population studied or by other phases of procedure which might not be apparent from a casual examination of the reports.

The reviewer may also test reported conclusions by considering their compatibility with general educational theory, with his experience in school affairs, and with logic or common sense. For example, the conclusion that drill in the fundamentals in arithmetic is effective in engendering skill in calculation, may be regarded as dependable since it is in harmony with the general principle that practice increases skill. This conclusion is also compatible with common sense. In case a reported conclusion does not appear reasonable or is not compatible with the reviewer's experience he is justified in being suspicious of the research. It, of course, does not follow that the reported conclusion is not dependable, and hence the apparent reasonableness of a conclusion cannot be considered a final criterion.

Taking notes on references. In taking notes on a reference, the reviewer should be guided by his problem. Rereading will be necessary if items pertinent to his problem are omitted. There is also a waste of time if any large amount of irrelevant information is included in the notes. The reviewer should, therefore, have clearly in mind the items of information to be looked for in each reference. It is frequently desirable to prepare a check list, or a data sheet to use as a guide in the reading of reports of educational research. For example, let us assume that the nature of the problem is such that most of the previous research may be expected to be of the experimental type. Then a check list containing the following items will be useful:

1. Reference
2. Experimental factor or factors
3. Dependent variable
4. School subject
5. Population, size of groups and their grade placement
6. Representativeness of population
7. Equivalence of experimental and control groups
8. Control of non-experimental factors
9. Duration of experiment
10. Tests used for measuring dependent variable
11. Differences in gains, or other measures of the relative achievement of the groups
12. Conclusions and generalizations
13. Evaluation

If many references are to be read, it is helpful to prepare mimeographed data sheets.¹ These sheets can be divided by horizontal and vertical lines into spaces for writing notes relative to the items to be noted. These spaces may be labeled with appropriate headings, but it is simpler to number them to correspond with the items of the check list. It should not be inferred, however, that a data sheet or check list can be used in a purely routine fashion. The reviewer will often encounter, in a report of research, information which is relevant to his problem, but which is not referred to by the items of the check list or the headings of the data sheet. When this occurs, notes pertaining to such information should be recorded.

The reviewer should strive to be accurate in note-taking. The notes should be checked against the report of research before going on to another study. Where quotations are included, this fact should be indicated by means of quotation marks. This precaution is one which may be instrumental in preventing the use of quoted material as one's own—unintentional plagiarism, but plagiarism nevertheless. In addition to indicating whether or not information is quoted, it is desirable to note the page on which the information is to be found. This

¹ The Alexander Universal Bibliography Card, obtainable from the Bureau of Publications, Teachers College, Columbia University, is a useful record form when only brief notes are being made.

should always be done for direct quotations and usually for the more important notes that are not quotations.

Organizing the research relating to a problem or topic. After the researches relating to the problem or topic have been reviewed they should be classified so as to bring together those pertaining to the same phase or that are similar in other respects. Usually the definition of the problem or topic will suggest captions for this classification. For example, in summarizing the research relating to the teaching of arithmetic the present writers employed the following major divisions: ¹

- I. Methods of teaching and learning the fundamentals
- II. Methods of drill in the fundamentals
- III. Methods of teaching pupils to solve verbal problems
- IV. Methods of diagnosis and remedial treatment
- V. Methods of teaching the reading of arithmetical subject-matter
- VI. Motivation of learning activity in arithmetic

Usually the particular outline evolved is not a matter of paramount importance. Frequently, there may be two or more organizations that will be effective as a plan for summarizing the research, but when the number of studies is large the formulation of an outline, such as that just illustrated, is an essential step. The effectiveness of a summary depends upon its organization. Hence, the problem and the several researches should be examined carefully in an effort to arrive at an effective plan of organization.

How much description of the researches to include in a summary. A troublesome problem in preparing a summary is to determine how much description of the researches to include. No general rule can be stated. If the number of researches summarized is large and each is described in detail, a reader is likely to wish for a summary of the summary. The criticism implied in such a wish may be partially avoided by separating the description of the researches from their evaluation. But a

¹ Monroe, W. S., and Engelhart, M. D. "A Critical Summary of Research Relating to the Teaching of Arithmetic," *University of Illinois Bulletin*, Vol. 29, No. 5, *Bureau of Educational Research Bulletin*, No. 58. Urbana, Illinois: University of Illinois, 1931. 115 pp.

reviewer should attempt to restrict his descriptions to essential items. When the research is of a conventional type, the description may be restricted to items that are indicative of the dependability of the findings. It is seldom desirable to report the findings in detail. Details of the treatment of the data should not be included unless they are essential for the evaluation. Sometimes the description of several similar researches may be combined and attention directed to points of similarity and points of difference. In many cases an abbreviated description will be satisfactory and in general a reviewer should endeavor to reduce the descriptions of the researches to a minimum, but no essential items should be omitted.

Evaluative statements important. The reviewer's evaluation should be included in the summary. Sometimes it may be sufficient to indicate the evaluation by such phrases as "carefully conducted experiment," "not convincing," "open to rather serious criticism," and "critically reported." In the case of major studies, it is desirable to point out the reasons for the evaluation. In general, a study should not be referred to in a summary without some indication of the reviewer's evaluation. Failure to conform to this rule is likely to contribute to the perpetuation of error and triviality.

Synthesis of dependable findings. A summary in which the several evaluated findings are reported in a serial order is not satisfying, especially when more than three or four studies relating to a particular problem are being considered. Such a summary suggests the notes taken as the studies were read and leaves to the reader the task of synthesizing the several findings into a composite conclusion. Unfortunately a large proportion of the available summaries are essentially nothing more than classified annotated bibliographies.

The description of the researches and the presentation of their findings should culminate in a synthesized statement of the conclusions justified by the researches as a group. This synthesis should be stated with precision and the dependability of the general conclusions should be made clear. Unless

the researches reviewed are so fragmentary and so unrelated that a synthesis is not feasible, a summary that does not even-tuate in a generalization may appropriately be called incom-plete.

ILLUSTRATIVE SUMMARIES

The following references are not cited as model summaries, but they are probably representative of the more scholarly writings of this type.

BROWNELL, W. A. "The Techniques of Research Employed in Arithmetic," Twenty-Ninth Yearbook of the National Society for the Study of Education. Bloomington, Illinois: Public School Publishing Com-pany, 1930, pp. 415-43.

BUSWELL, G. T., and JUDD, C. H. "Summary of Educational Investigations Relating to Arithmetic," *Supplementary Educational Monographs*, No. 27. Chicago: University of Chicago, 1925. 212 pp.

COREY, S. M. "The Present State of Ignorance about Factors Effecting Teacher Success," *Educational Administration and Supervision*, 18: 481-90, October, 1932.

CROOKS, A. D. "Marks and Marking Systems: A Digest," *Journal of Educational Research*, 27: 259-72, December, 1933.

ENGELHART, M. D. "Techniques Used in Securing Equivalent Groups," *Journal of Educational Research*, 22: 103-09, September, 1930.

FRYER, DOUGLAS. *The Measurement of Interests*. New York: Henry Holt and Company, 1931. 488 pp.

HILLIARD, G. H. "Probable Types of Difficulties Underlying Low Scores in Comprehension Tests," *University of Iowa Studies, Studies in Educa-tion*, Vol. 2, No. 6. Iowa City: University of Iowa, 1924, pp. 13-36.

HUDELSON, EARL. "Class-Size Opinions, Evidence, and Policies in Sec-ondary Schools," *North Central Association Quarterly*, 4: 196-208, September, 1929.

KNUDSEN, C. W. "Psychology and Methods in the High School and College—Social Studies," *Review of Educational Research*, 4: 462-65, December, 1934.

LEE, J. M., and SYMONDS, P. M. "New-Type of Objective Tests: A Sum-mary of Recent Investigations," *Journal of Educational Psychology*, 24: 21-38, January, 1933.

LEONARD, J. P. "Psychology and Methods in the High School and Col-

- lege—English Language, Reading, and Literature," *Review of Educational Research*, 4: 449-61, December, 1934.
- LYMAN, R. L. "Summary of Investigations Relating to Grammar, Language, and Composition." *Supplementary Educational Monographs*, No. 36. Chicago: University of Chicago, 1929. 302 pp.
- MONROE, W. S., and ENGELHART, M. D. "Stimulating Learning Activity," *University of Illinois Bulletin*, Vol. 28, No. 1, *Bureau of Educational Research Bulletin*, No. 51. Urbana: University of Illinois, 1930, pp. 42-57.
- NATIONAL EDUCATION ASSOCIATION, RESEARCH DIVISION. "Contributions of Research to Curriculum Building," *Research Bulletin*, Vol. 3, Nos. 4 and 5. Washington: National Education Association, 1925, pp. 125-61.
- POWERS, S. R. "Psychology and Methods in the High School and College—Science," *Review of Educational Research*, 4: 473-78, December, 1934.
- ROCK, R. T., JR. "A Critical Study of Current Practices in Ability Grouping," *Catholic University of America, Educational Research Bulletin*, Vol. 4, Nos. 5 and 6. Washington: Catholic Education Press, 1929. 132 pp.
- SYMONDS, P. M. "Methods of Investigation of Study Habits," *School and Society*, 24: 145-52, July 31, 1926.
- YANKEY, J. V., and ANDERSON, P. L. "A Review of the Literature on the Factors Conditioning Teaching Success," *Educational Administration and Supervision*, 19: 511-20, October, 1933.

CHAPTER XIV

PROGRESS TOWARD A SCIENCE OF EDUCATION

The development of a scientific attitude in education. A significant phase of our progress toward a science of education¹ is found in the growth of a scientific attitude in education. When Rice reported his study of spelling at the meeting of the Department of Superintendence in 1897, the audience was distinctly hostile to the implied thesis that the outcomes of spelling instruction could be measured by administering a test.² The

¹ Some readers will doubtless raise the question of the possibility of a science of education. This question is not new and it is interesting that many writers answered it in the affirmative before 1900. In the following list, Royce is the only one who does not give an affirmative answer.

Payne, J. "Science of Education," *Barnard's American Journal of Education*, 26: 465-68, 1876.

Allen, Jerome. "Have We a Science of Education?" *Education*, 2: 284-90, January, 1882.

Bain, A. *Education as a Science*. New York: Appleton Company, 1884. 453 pp.

Payne, W. H. *Contributions to the Science of Education*. New York: American Book Company, 1886. 358 pp.

Royce, J. "Is there a Science of Education?" *Educational Review*, 1: 15-25, January, 1891.

Scripture, E. W. "Education as a Science," *Pedagogical Seminary*, 2: 111-14, 1892.

Findlay, J. J. "The Scope of a Science of Education," *Educational Review*, 14: 236-47, October, 1897.

For a vigorous defense of the thesis that a science of education is possible, see Phillips, D. E. "What 'Is Scientific?'" *Journal of Educational Psychology*, 23: 299-308, April, 1932. In this article the suggestion is made that if mathematicians and workers in the field of the more exact sciences wish to take issue with the thesis that a science of education is possible, different levels of scientific endeavor might be recognized.

The interested reader will find the following reference helpful:

Demiashkevich, M. J. "The Science of Education," *Phi Delta Kappan*, 14: 184-86, April, 1932.

² This study concerns the relation between the minutes per day devoted to the teaching of spelling and the spelling ability of the pupils. See page 271. For the original report, see Rice, J. M. "The Futility of the Spelling Grind," *The Forum*, 23: 163-72, 409-19; April, June, 1897. These articles also appear as Chapters V and VI in Rice, J. M. *Scientific Measurement in Education*. New

reaction of this audience is indicative of the prevailing attitude at the close of the nineteenth century. Galton,¹ James,² Hall,³ Cattell,⁴ and other psychologists were attempting to apply the methods of science, but their efforts had little direct influence upon the thinking of school administrators and teachers.

During the first decade of the twentieth century, we find indications of a change in attitude. There was a marked increase in the number of experimental studies of the transfer of training.⁵ In 1904 Superintendent Maxwell of New York City included in his annual report an age-grade study of the elementary schools of that city.⁶ The appearance of this report appears to have stimulated interest in the questions of retardation and elimination. Within a period of less than ten years, a number of elaborate studies were made, of which Thorndike's study, "The Elimination of Pupils from School,"⁷ in 1907

York: Hinds, Noble, and Eldredge, 1912. Rice's account of the reception of his report is given on pages 17-18 of this reference.

¹ Galton, Francis. *Hereditary Genius: An Inquiry into Its Laws and Consequences*. London: The Macmillan Company, 1914. 379 pp. (First Edition, 1869.)

Galton, Francis. *Inquiries into Human Faculty and Its Development*. London: The Macmillan Company, 1883. 387 pp.

² James, William. *Principles of Psychology*, Vol. 1. New York: Henry Holt and Company, 1890, pp. 666-68, footnote.

³ Hall, G. S. "The Contents of Children's Minds on Entering School," *Pedagogical Seminary*, 1: 138-73, 1891.

Hall, G. S. *Life and Confessions of a Psychologist*. New York: D. Appleton and Company, 1923. 623 pp. The student will find this autobiography an excellent source of information with respect to Hall's work.

⁴ Cattell, J. McK. "Mental Tests and Measurements," *Mind*, 15: 373-80, July, 1890.

Cattell, J. McK., and Farrand, Livingston. "Physical and Mental Measurements of the Students of Columbia University," *Psychological Review*, 3: 618-48, November, 1896.

⁵ Rugg, H. O. *The Experimental Determination of Mental Discipline in School Studies*. Baltimore: Warwick and York, Inc., 1916. 132 pp.

Rugg gives an analytical summary of twenty-nine studies. Three appeared before 1900, six during the next five years, and twenty during the period 1906-1916. It is significant that only one study of transfer under school conditions was made before 1906 whereas nine such studies were made during the ten years following.

⁶ Maxwell, W. H. *Sixth Annual Report of the City Superintendent of Schools*. New York, 1904, pp. 42-49.

⁷ Thorndike, E. L. "The Elimination of Pupils from School," *U. S. Bureau of Education Bulletin*, No. 4. Washington: Government Printing Office, 1907. 63 pp.

appears to have been the first. It was concerned chiefly with elimination, but some attention was given to retardation and acceleration. A couple of years later, 1909, Ayres published a somewhat more comprehensive investigation under the title *Laggards in Our Schools*.¹ In 1911, Strayer published a study² that presented age-grade data for a number of city school systems, colleges, and universities. In the same year two other reports appeared, one³ of which dealt chiefly with the progress of pupils, rather than with age-grade conditions, and the other⁴ with retardation. Meyer's study of teachers' marks⁵ reported in 1908 was followed by investigations by Dearborn,⁶ Starch and Elliott,⁷ and Kelly.⁸ The First Yearbook of the National Society of College Teachers of Education, published in 1911, was devoted to the subject "Research within the Field of Education, Its Organization and Encouragement."⁹ Binet, who had been working with psychological tests for a number of years, devised and published in collabora-

¹ Ayres, L. P. *Laggards in Our Schools*. New York: Charities Publication Committee, 1909. 236 pp.

² Strayer, G. D. "Age and Grade Census of Schools and Colleges," *U. S. Bureau of Education Bulletin*, No. 5. Washington: Government Printing Office, 1911. 144 pp.

³ Keyes, C. H. "Progress through the Grades of City Schools," *Teachers College, Columbia University Contributions to Education*, No. 42. New York: Bureau of Publications, Columbia University, 1911. 79 pp.

⁴ Blan, L. B. "A Special Study of the Incidence of Retardation," *Teachers College, Columbia University Contributions to Education*, No. 40. New York: Bureau of Publications, Columbia University, 1911. 111 pp.

⁵ Meyer, Max. "The Grading of Students," *Science*, 28: 243-52, 1908.

⁶ Dearborn, W. F. "The Relative Standing of Pupils in High School and in the University," *University of Wisconsin Bulletin*, No. 312, 1909. 44 pp.

⁷ Starch, Daniel, and Elliott, E. C. "Reliability of Grading High School Work in English," *School Review*, 20: 442-57, September, 1912.

⁸ Kelly, F. J. "Teachers' Marks," *Teachers College, Columbia University Contributions to Education*, No. 66. New York: Bureau of Publications, Columbia University, 1914. 139 pp.

⁹ "Research within the Field of Education, Its Organization and Encouragement," *School Review Monographs*, No. 1. Chicago: University of Chicago Press, 1911. 71 pp.

The major portion of this volume consists of four papers:

Cubberley, E. P. "Fundamental Administrative Problems."

Dearborn, W. F. "Experimental Education."

Monroe, Paul. "Coöperative Research in Education."

Thorndike, E. L. "Quantitative Investigations in Education: with Special Reference to Coöperation within This Association."

tion with Simon the Binet-Simon General Intelligence Scale ¹ in 1905. This scale was introduced in the United States by Goddard who published a standardization for American children in 1910. Other revisions were published in 1912 by Kuhlmann and by Terman and his coworkers.²

The work of Rice served as a stimulus to Thorndike, Stone, Courtis, and others. Thorndike published the first book dealing directly with mental measurement ³ in 1904. Courtis, who had coöperated with Stone ⁴ by administering his tests, constructed a group of arithmetical tests designated as Series A which were made available for use in September, 1909. Thorndike's Handwriting Scale was published in March, 1910, and the Hillegas English Composition Scale in 1912.

In an address before the Harvard Teachers Association in March, 1912, Ayres commented on the change in attitude since Rice made his report at the meeting of the Department of Superintendence in 1897.

Last week, in the City of St. Louis, that same association of school superintendents, again assembled in convention, devoted forty-eight addresses and discussions to tests and measurements of educational efficiency. The basal proposition underlying this entire mass of discussion was that the effectiveness of the school, the methods, and the teachers must be measured in terms of the results secured.

This change represents no passing fad or temporary whim. It is permanent, significant, and fundamental. It means that a transformation has taken place in what we think as well as in what we do in education.⁵

¹ Binet, A., and Simon, T. "Méthodes Nouvelles pour le Diagnostic du Niveau Intellectuel des Anormaux," *L'Année Psychologique*, 11: 191-244, 1905.

² For a brief historical account of intelligence testing, see Pintner, Rudolph. *Intelligence Testing*. New York: Henry Holt and Company, 1923, Chapters I, II, and III.

³ Thorndike, E. L. *An Introduction to the Theory of Mental and Social Measurement*. New York: Teachers College, Columbia University, 1904. 277 pp. (Revised edition, 1913.)

⁴ Stone, C. W. "Arithmetical Abilities and Some Factors Determining Them," *Teachers College, Columbia University Contributions to Education*, No. 10. New York: Bureau of Publications, Teachers College, Columbia University, 1908. 101 pp.

⁵ Ayres, L. P. "Measuring Educational Processes through Educational Results," *School Review*, 20: 300-01, May, 1912.

The final stand of those in opposition to the use of educational tests and the interpretation of the resulting average scores by comparison with established norms was made on the proposition that such practices would result in a standardization of education that would be opposed to instructional efficiency. At the meeting of the Department of Superintendence in February, 1915, the National Council of Education planned its program so as "to give full hearing to those who are skeptical about the desirability of standardization, tests, measurements, and other exact forms of evaluating school work." The program of the first session of this organization bore the general title, "Standardization, Wise and Otherwise." Opportunity was given to some of the outstanding opponents of measurement to present their case. Referring to this meeting ten years later, C. H. Judd said:

There are many here who will recall the meeting of the National Council in 1915 when the forces of conservatism gathered for a last stand and a battle was fought to determine whether measurement of mental and moral traits was to be recognized as permissible.

There can be no doubt as we look back on that council meeting that one of the revolutions in American education was accomplished by that discussion. Since that day, tests and measures have gone quietly on their way as conquerors should. Tests and measures are to be found in every progressive school in the land. The victory of 1915 slowly prepared during the preceding twenty years was decisive.¹

The present attitude toward educational research is one of confidence in its possibilities. Superintendents look to educational research for the construction of the curriculum, for the determination of the relative merits of various practices, for the evaluation of textbooks, and for the answers to many other questions in the field of education. Teachers expect educational research to tell them the relative merits of various methods and devices of teaching. Many schoolmen are apparently looking forward to the time when most, if not all questions which are

¹ Judd, C. H. "The Curriculum: A Paramount Issue," *Addresses and Proceedings of the National Education Association*, Vol. 63. Washington: National Education Association, 1925, pp. 806-07.

now confusing and hence are a constant source of irritation because they must be thought about, will have been answered by educational research and answered conclusively so that it will no longer be necessary for one's peace of mind to be disturbed by trying to think about them.

Development of techniques for educational research. A second phase of our progress toward a science of education is the development of research techniques. Before the publication of Rugg's text ¹ in 1917, our principal sources in regard to statistical methods were Thorndike, *Theory of Mental and Social Measurement*, and Yule, *An Introduction to the Theory of Statistics*. The citations in Chapters IV, X, and XI to recent writings in this field indicate the contributions to statistical methods and the development of instruments for measuring human traits and abilities and of other research techniques is reflected in the discussion of other chapters of this volume. It is apparent that marked progress has been made in this phase of the development of a science of education.

Research activities in the field of education. It is apparent that there has been much activity in the field of educational research during recent years, but the mention of certain facts will serve to make the picture of this phase of our progress toward a science of education more definite. The report of the New York School Inquiry, 1911-1912, included the recommendation that a "Bureau of Investigation and Appraisal" be established. As a result of this recommendation, a Division of Reference and Research was established in 1913. Similar departments were organized in other cities: Baltimore, 1912; Rochester, N. Y., 1913; New Orleans, 1913; Boston, 1914; Kansas City, Missouri, 1914; Detroit, 1914; Schenectady, N. Y., 1914; Oakland, California, 1914. The establishment of departments of educational research in educational institutions was due largely to the suggestion of S. A. Courtis, who had developed the idea of comparative testing advocated by Rice.

¹ Rugg, H. O. *Statistical Methods Applied to Education*. Boston: Houghton Mifflin Company, 1917. 410 pp.

At first, Courtis directly solicited the coöperation of superintendents and teachers in standardizing the tests he devised. As the interest in the testing movement grew, he foresaw the desirability of having centers in each state for distributing the tests, receiving and compiling the scores obtained, and inquiring into conditions that appeared unusual. Such centers were established at the University of Oklahoma, 1913; Indiana University, 1914; Kansas State Normal School, Emporia, 1914; University of Iowa, 1914; University of Minnesota, 1915. The first state bureau was the Division of Educational Tests and Measurements of the Wisconsin State Departments of Public Instruction, organized in 1916. The Iowa Child Welfare Research Station was authorized by the Iowa General Assembly in 1917, and the Bureau of Educational Research of the University of Illinois was established by action of the Board of Trustees in 1918.

Beginning about 1920 there is evident a definite effort to increase the production of educational research. For example, in 1921 the directors of the Commonwealth Fund appropriated \$100,000 a year for a period of five years to be used in subsidizing investigations by individuals and organizations. The attitude of the committee administering the fund is indicated in the following paragraph from a statement issued by the secretary of the committee at the end of the first year:

The Educational Research Committee believes that there should be many more appeals for subventions than have thus far come to it and that requests should be made by a much wider range of institutions. Indeed the conditions of the grant and the policy of the committee are so flexible that any first-class project which can be clearly defined and budgeted is likely to receive favorable consideration. The committee meets three times a year, in the autumn, in the early spring, and in the early summer.¹

One of the features of a volume published in 1926 was a plea for research by teachers.² In connection with such appeals,

¹ Editorial. *Elementary School Journal*, 22: 404, February, 1922.

² Buckingham, B. R. *Research for Teachers*. New York: Silver, Burdett and Company, 1926. 386 pp.

teachers were told participation in experimentation and other types of educational research is relatively simple and requires little if any special training. For example, in the book just referred to, it is asserted that "it is by no means necessary that you should set up formal experiments involving control groups in order to serve the cause of education as a research worker."¹ A similar assertion was made in an editorial announcement in the *English Journal* for February, 1923, p. 138. The editor proposed an experiment to determine the relative merits of two instructional procedures. After explaining the plan of the inquiry and soliciting the coöperation of teachers, the writer stated:

No technical training in the use of measurements will be necessary, and there will be no great additions to the teacher's out-of-class labors. Only the collection of a few samples of his own pupils' compositions and fairly close adherence to definite teaching policies in two classes—these will be the total burden of each co-operator.

An indication of the amount of research activity is afforded by the number of doctors' degrees conferred in education. Table XVI shows the number by years from 1918-1932. From 1918 to 1922 inclusive, the average number per year was 55, 68 for 1922 being the largest. In 1923, the number rose to 94, and in 1926 to 181, and in 1932 to 337. Further evidence of the attainment of quantity production in educational research is afforded by the annual bibliographies of educational research published by the United States Office of Education. For example, the bibliography published in 1928 and referring to research reported in 1926-1927 included 1540 titles. The bibliography published in 1931 lists 4651 studies reported in the years 1929 and 1930. In commenting on this list of studies, an editorial in the January, 1932, issue of *School Life* asserts that on a conservative basis the 4651 studies listed represent a total expenditure of time and money of not less than ten million dollars.

¹ *Ibid.*, p. 377.

TABLE XVI. NUMBER OF DOCTORS'
DEGREES IN EDUCATION

YEAR	NUMBER
1918	53
1919	50
1920	61
1921	43
1922	68
1923	93
1924	110
1925	137
1926	181
1927	189
1928	189
1929	218
1930	265
1931	308
1932	337
Total	2302

Contributions to a science of education. The most significant measure of our progress toward a science of education is to be found in the accumulation of contributions to it. The attainment of "quantity production" in research activities suggests a large accumulation of contributions. It is apparent, however, that many of the investigations commonly designated as educational research do not contribute to a science of education, at least, directly. A more conservative picture ¹ of the contributions is furnished by the bibliographies of the summaries published in the *Review of Educational Research*, 1931-1933. During this three-year period, the fifteen numbers of this journal were planned to cover the entire field of educational research. Hence, the studies listed may be thought of as representing the "cream" up to the time the several summaries were prepared. The general titles of the fifteen issues of the

¹ An interesting picture of the growth of educational research since 1890 has been contributed by Franke and Davis who classified 2837 articles from 13 periodicals appearing during this period.

Franke, P. R., and Davis, R. A. "Changing Tendencies in Educational Research," *Journal of Educational Research*, 23: 133-45, February, 1931.

journal and the numbers of items in the bibliographies are as follows:

1931	
The Curriculum.....	303
Teacher Personnel.....	458
School Organization.....	300
Special Methods in the Elementary	
School.....	438
Psychology in the School Subjects...	884
1932	
Special Methods on High School Level	276
Finance and Business Administration	449
Tests of Personality and Character ...	282
Tests of Intelligence and Aptitude....	450
School Buildings, Grounds, Equipment,	
Apparatus, and Supplies.....	476
1933	
Educational Tests and Uses.....	467
Mental and Physical Development....	433
Pupil Personnel, Guidance, and Coun-	
seling.....	793
Psychology of Learning, General	
Methods of Teaching, and Super-	
vision.....	457
The Legal Basis of Education.....	398
Total.....	6864

There is some duplication among the several bibliographies, but the additions due to duplication are probably much less than the number of unpublished studies not listed and the studies omitted because they had been dealt with in published summaries. Hence, the above total is probably a conservative measure of the "better" educational research completed up to the time that these summaries were prepared. The totals for 1934 and 1935, the first two years of the second cycle, are materially greater than those for 1931 and 1932. This increase is due in part to a more thorough canvass of educational literature for reports of researches and to the inclusion of a larger number of unpublished studies, but it is likely that the production rate has increased.

The status of the science of education. An estimate of the status of the science of education should recognize the nature

of its content. A science of education will include factual statements of the characteristics of defined populations of children such as age groups, gifted children, and the like, but these items will be incidental to statements of relationships and laws. A large section of the science of education will deal with the identification of the factors that affect human learning under school conditions,¹ and with the relationship between these factors and pupil achievement. Educational tests and other instruments for measuring human abilities and traits will receive attention, but the treatment will not be merely a description of their structure and directions for their application. At present there is much ignorance concerning what we measure. In a science of education there will be more adequate definition of the various abilities and traits that we attempt to measure. It is probable that standardized units of measurement will be defined.

It should be emphasized that statements of relationship, which will be prominent in a science of education, are generalizations. A sample of a gas has the same characteristics as any other sample, and hence, the relationship between volume, pressure, and temperature determined from it are applicable to any mass of the gas. Human beings are characterized by individual differences and the relationships we seek in the field of education are for averages within given populations. An experimental study of the relative effect of two instructional procedures is designed to reveal the difference between the average effects for a particular population.² The findings for another population may be different and hence in a science of education there will be statements of relationships for various typical populations.

The picture of the status of the science of education suggested

¹ For an indication of what is involved, see pages 278-89. The interested reader should consult also Courtis, S. A. "Factors Conditioning Growth," *Papers of the Michigan Academy of Science, Arts and Letters*, 10: 349-67, 1928.

² In theory the population might be an individual pupil with certain characteristics, but in such a case the findings would have such limited application that the value would be negligible.

by the bibliographies in the first three volumes of the *Review of Educational Research* is too optimistic. Examination of the summaries reveals that many of the authors were not very critical in compiling their bibliographies. Furthermore, many of the dependable findings can be regarded as only minor or incidental contributions to a science of education. We have accurate information concerning the distribution of teachers' marks in certain schools, the scores made upon certain tests by certain groups of pupils under certain conditions, the salaries of certain groups of teachers, the vocabularies of certain textbooks, the historical and geographical allusions in certain newspapers and periodicals, the eye-movements of certain readers, the age-grade status of certain school populations, the time allotted to the different studies in the elementary school curriculum, and the like. Such information is useful, but a compilation of it does not constitute a science of education. It is possible, on the other hand, to point to a number of studies whose conclusions are of a different sort and rightfully deserve to be recognized as contributions to a science of education. For example, research has given us valuable information regarding the relation of eye movements to the reading process. It has also contributed to the formulation of laws of learning. We now know a great deal concerning the predictive value of many measures. A critical survey of educational research would doubtless reveal several hundred studies that should be regarded as significant contributions to a science of education.¹

The status of the science of education may be viewed also with reference to the character of the problems studied. Freeman² has pointed out that a large share of educational research has been devoted to disproving theories and hypotheses. It is,

¹ Since the nature of a science of education suggests that controlled experimentation is a very fruitful type of research, Chapter IX should be reviewed in connection with the study of the present topic. The concluding pages are especially applicable.

² Freeman, F. N. "The Contributions of Science to Education," *School and Society*, 30: 107-12, July 27, 1929.

of course, important to have hypotheses subjected to experimental verification, but when research shows that a hypothesis is not tenable, only a negative contribution to the science of education has been made. Positive contributions are necessary for building up a science of education.

In Chapter II attention was called to the fundamental problems of education. Relatively few research workers are directing their time and energies to these problems. Survey investigations reporting the status of current practices and conditions make only incidental contributions to a science. In the field of educational measurements many instruments have been constructed but relatively few attempts have been made to identify and define the ability or trait whose measurement is desired. We have many scales for measuring teaching efficiency but there is no authoritative and precise definition of what is designated by this term. Factual items such as coefficients of correlation between certain measures are not useful in building up a science of education when they refer to undefined populations or to measures whose validity is unknown. Many experimental studies have been based upon populations that are not representative or are so small that generalization is hazardous.

A third indication of the status of the science of education is the relatively small number of experimental findings that have been verified. Although critical evaluation of reported studies will usually lead to a fairly trustworthy estimate of the dependability of the findings, the ultimate test of a generalization is verification by a repetition of the investigation. In chemistry, physics, and other scientific fields much importance is attached to the corroboration of reported findings, but in the field of education investigators interested in the same problem have seldom employed sufficiently similar techniques to justify precise comparison of their results. As a rule, educational research workers have been much more interested in a new investigation or in applying an improved technique than in repeating a study for the purpose of testing the dependability of reported findings.

The large investment in educational research during the past fifteen years has been largely devoted to isolated studies rather than to coördinated inquiries. In other words, our research activities have not fitted into a general program. There have been a few relatively comprehensive programs such as the Terman Study of Gifted Children, the Educational Finance Inquiry, the Modern Foreign Language Investigation, and the Charters' Study of Motion Pictures and Youth but for the most part the studies bearing upon a general problem or topic are so lacking in coördination that the findings can only be described as fragmentary and a synthesis of them is so lacking in completeness that generalization is not justified.¹

A fifth indication of the status of the science of education is the number of persons who have become definitely interested in educational research. The great volume of graduate theses in education and other research writings suggests that the number of such persons is very large. The same indication is given by the number of elections to Phi Delta Kappa, an honorary society emphasizing interest in research in selecting its members, which now (1933) totals more than 15,000. The actual number of persons genuinely interested in educational research is, however, probably not very large. Brownell² states that out of a total of seventy unpublished masters' theses included by Gray in his summary of research relating to the field of reading up to 1929, only portions of twelve were later published. This condition, which is probably representative of masters' theses in education, indicates either that the other fifty-eight theses were not considered worthy of a published account or that the authors did not have sufficient interest in their work to prepare an account for publication. A number of persons have demonstrated a persistent interest in educational research by continuing their inquiries and publishing their findings, but the total of such persons is small in com-

¹ Hartmann, G. W. "Laissez Faire versus Planning in Educational Research," *School and Society*, 39: 600-03, May 12, 1934.

² Brownell, W. A. "The Growth and Nature of Research Interest in Arithmetic and Reading," *Journal of Educational Research*, 26: 440, February, 1933.

parison with the number whose names appear once, or at most twice in comprehensive bibliographies.

A number of writers have expressed judgments relative to the status of the science of education. The following are perhaps typical.

The chief service of contributions in the field of educational research up to the present time has undoubtedly been in pointing out problems and methods of approach.¹

We must use greater care to make certain that the conclusions we state in our reports follow logically from the data presented. Too many reports state conclusions that are not fully supported by the research data included in them.²

Nevertheless, I cannot evade the conviction that, relatively speaking, the published research in education is, on the whole, inferior in quality, and more especially inferior in ultimate significance, to the published research in other branches of scientific endeavor. Too many contributions seem essentially futile. After you read them, you feel like saying: "Well, suppose it is true; what of it?"³

Writing in 1928, Courtis described the status of the science of education as that of "biased observation and uncritical acceptance of assumptions."⁴ In support of this evaluation he calls attention to our ignorance concerning what our instruments measure and asserts that "we have not yet identified what it is that is measured by any test."

After an extensive inquiry into the facilities for educational research in public school systems and an examination of a large number of reports of research, Zeigel states that "the bureaus of research in city systems are concerned primarily with the mere compilation of facts and statistics and consequently do not meet the prerequisite qualities of educational research pro-

¹ Theisen, W. W. "Recent Progress in Educational Research," *Journal of Educational Research*, 8: 314, November, 1923.

² Trabue, M. R. "Educational Research in 1925," *Journal of Educational Research*, 13: 344, May, 1926.

³ Whipple, G. M. "The Improvement of Educational Research," *School and Society*, 26: 251, August 27, 1927.

⁴ Courtis, S. A. "Education—A Pseudo-Science," *Journal of Educational Research*, 17: 131, February, 1928.

mulgated by leaders in education.”¹ In another place he states “the upshot of the facts presented is that the total extent and the quality of the research carried on within school systems is not highly commendable.”²

These evaluations of educational research may seem to be unduly critical and it may be pointed out that several of them are not recent.³ But it must be admitted that critical contemplation of the accomplishments of educational research does not lead to a very high estimate of the status of the science of education.⁴ The accomplishments are not negligible and in view of the short period during which there has been persistent effort to build up a science of education and the relatively slow development of the physical sciences,⁵ we may with some justification point to them with pride. But it seems a fair statement to say that we are just beginning to be aware of the nature of the task before us. Fundamental problems are beginning to receive the attention of an increasing number of workers with time and resources for research and the faith they exhibit is a significant indication.

The contributions of educational research to educational practice.⁶ An appraisal of the research movement in education would be incomplete without directing attention to certain contributions to educational practice. The direct application

¹ Zeigel, W. H., Jr. “Research in Secondary Schools,” United States Department of the Interior, Office of Education, Bulletin No. 17, National Survey of Secondary Education Monograph, No. 15. Washington: United States Government Printing Office, 1933, p. 66.

² *Ibid.*, p. 71.

³ Several of the more recent writings on this topic have been as critical as the statements just quoted. For example, see

Symonds, P. M. “Common Faults in Graduate Research in Education,” *Journal of Educational Research*, 27: 481-92, March, 1934.

⁴ The reader who has not studied the preceding chapters, especially VIII, IX, and X should read them in this connection.

⁵ Curtis has made an interesting comparison of progress in the science of education with the development in other scientific fields. Curtis, S. A. “The Construction of Measuring Instruments in the Field of Education,” *Scientific Monthly*, 21: 260-90, September, 1925.

⁶ For a more extended discussion, see Monroe, W. S. “Service of Educational Research to School Administrators,” *American School Board Journal*, 70: 37-39, 122, 125, April, 1925.

of research findings has resulted in a number of changes in our schools, some of them destined to have far reaching effects.¹ Another effect of the research movement is evidenced by the utilization of statistical techniques,² educational tests, and other instruments in the study of practical school problems.³ A third contribution, although somewhat less tangible, is to be found in the analysis and definition of problems studied by research workers. Many practical problems which appear relatively simple to the uninitiated have been revealed as highly complex.⁴

Progress toward a science of education retarded by a false concept of research. Many graduate students and other persons who have been introduced to educational research during recent years have gained the impression that by carrying through certain procedures, mostly routine in character, answers to educational problems would be revealed. This is a false concept of educational research. It is true that in certain types of investigations there is much routine work, but even the routine procedures of educational research are based upon assumptions that must receive attention in interpreting the findings. Furthermore, educational data involve errors both of measurement and of validity which must receive attention. The movement for quantity production served to "sell" the idea that educational research is possible and desirable, but by trying to make it appear simple, the leaders supporting this movement contributed to building up the false concept just noted. The requirement of a thesis for a graduate degree and the custom of designating as research the activity of satisfying this requirement, has doubtless contributed to the same end. The engendering of this false concept of research in education has been contributed

¹ For a discussion of the ways in which research has modified educational practice, see Judd, C. H. "Educational Research and the American School Program," *Educational Record*, 4: 165-77, October, 1923.

² See Douglass, H. R. "The Contribution of Statistical Method to Education," *School and Society*, 35: 815-24, June 18, 1932.

³ For an account of the application of research techniques in the field of administration, see Strayer, G. D. "The Scientific Approach to the Problems of Educational Administration," *School and Society*, 24: 685-95, December 4, 1926.

⁴ For an elaboration of this point, see pages 289-93 and 424.

to by many persons occupying positions of distinction but who have been inadequately trained in statistical methods. Frequently, such persons have accepted uncritically findings that are not dependable and by giving publicity to them have caused their audiences to believe that educational research is more simple than it is.

Perhaps more important is the effect created by clever and fluent speakers and writers who decorate their arguments with uncritical citations of findings which are in agreement with their opinions. Such persons take advantage of the attitude toward research that has been built up, and create the impression of being scientific when they are not. The point was made in Chapter XII that questions which ask what should be, cannot be answered by means of objective methods. Many of the educational problems that now command our attention are of this type, and a writer or speaker, who gives the impression that his answer to such a question has been demonstrated by research, takes an unjustifiable advantage of his audience. Eventually audiences will detect such uncritical thinking. They may possibly conclude that such writers and speakers are lacking in sincerity. It is not unlikely that disillusionment will result in building up an indifferent, if not unfriendly, attitude toward educational research. This danger, which in the judgment of the present writers is a very real one, would be minimized if writers and speakers were more critical in their use of reported findings and were willing to have their assertions appear as a product of their own thinking.

The retarding influence of this false concept of educational research is difficult to estimate, but to one who has been reading reports of studies for more than twenty years it appears to have been considerable. However, the mistakes of the past are behind us, and it is now clear that there should be concerted effort to engender a more adequate understanding of what is involved in educational research. The genuine friends of educational research should recognize their responsibility. They should be critical not only in their own investigations but also in evaluat-

ing the work of others. They should seek an adequate understanding of research techniques and refrain from giving uncritical publicity to findings whose dependability is uncertain. Graduate students and other amateurs should not be encouraged to undertake studies for which they do not have adequate training. Unfortunately, the treatment of research techniques in our better texts is in some cases inadequate, and in a few it is misleading or even erroneous. This condition adds to the difficulty of attaining an adequate understanding, but persistent efforts on the part of the genuine friends of educational research will gradually overcome this handicap.

Crucial needs of educational research. The need for identifying and formulating the assumptions basic to a science of education is fundamental. Some attention has been given to this matter,¹ but we do not as yet have a comprehensive formulation. As assumptions are formulated, they should be checked against each other and against available research findings and the assumptions that appear reasonable should then be organized to serve as a basis for research directed toward building up a science of education.

It is frequently asserted that the techniques of educational research should be refined and probably the statement would be accepted by all competent persons, but it remains to inquire what refinements are needed. The difficulties encountered in educational research have been systematically noted in the preceding chapters and it will be sufficient to note here only the more important needs for improvement.

There is need for a more adequate understanding of statistical methods. Many research workers employ statistical techniques

¹ For illustrations see

Courtis, S. A. "The Factor Concept in Education," *School and Society*, 19: 413-23, April 12, 1924.

Bode, B. H. "Where Does One Go for Fundamental Assumptions in Education?" *Educational Administration and Supervision*, 14: 361-70, September, 1928.

Freeman, F. N. "Psychology as the Source of Fundamental Assumptions in Education," *Educational Administration and Supervision*, 371-77, September, 1928.

Hendrickson, Gordon, "Some Assumptions Involved in Personality Measurement," *Journal of Experimental Education*, 2: 243-49, March, 1934.

which they understand only imperfectly. This is especially true of correlation techniques and probable error formulae. The derivation of most statistical formulae is based upon assumptions in regard to the data to which they are to be applied and other assumptions ¹ are introduced in the use and interpretation of the statistics resulting from the application of such formulae. Without an understanding of the underlying assumptions, it is easy to misinterpret statistics and many instances of erroneous or misleading use of statistical techniques are to be found in our educational literature.

A second need relates to the refinement of the measuring instruments. Considerable attention has been given to improving the reliability of measures, but increasing the reliability of measuring instruments is not sufficient. Attention must be given also to their validity. The fact that measures to which we give different labels appear to overlap to a marked degree, indicates a need for precise definition of the measures yielded by educational tests and other instruments.

The educational research worker seldom enjoys the privilege of working with perfect data. Almost always they involve errors of measurement. Frequently they involve errors of validity. Sometimes the data as a group do not conform to the assumptions upon which the statistical techniques are based and when generalization is attempted it is necessary to inquire concerning the representativeness of the sample used. Hence, it is imperative that research workers ascertain the faults of the data with which they are working and make due allowance for these faults in interpreting their findings. We have made considerable progress in identifying the faults of data, but there is need for further inquiry, especially relative to the causes and magnitude of systematic errors of measurement and of validity. The matter of sampling also should be studied.

Refinement of controlled experimentation involves more precise definition of the experimental factor as well as more effective

¹ These assumptions are in addition to those referred to in the first paragraph of this section.

control of non-experimental factors. There is need also for maintaining a status of the non-experimental factors that is compatible with sound educational practice. Longer periods of experimentation and more representative groups of pupils will add greatly to the dependability of the findings.

In addition to effecting such refinements of the techniques of educational research, it is imperative from the point of view of building up a science of education that research workers direct their energies to the more fundamental problems. When we inquire into the problems that research workers have studied, it is found that relatively few of them have given attention to the fundamental problems of education noted in Chapter II. Much of the time and money devoted to educational research has been expended for survey investigations—many of them being trivial or having only local significance. Relatively little progress toward the science of education will be attained until research workers generally devote themselves to the basic problems of education when it is considered as a science.

In directing their efforts to the fundamental problems of education, research workers should engage in more long time investigations. This is especially needed in experimental studies in which a segment of achievement is the dependent variable. Few experiments have been continued more than a few months. In some cases the experimental period has been limited to a few weeks. It appears probable that in many cases such short time investigations do not reveal adequately some of the more important effects of controlled changes in the experimental factor. There is also need for more intensive studies. The Chicago Reading Studies illustrate what may be accomplished when a worker devotes himself to a limited field of investigation for a period of years.

The challenges of educational research. The fundamental problems are complex and difficult, but they constitute challenges to the student of education interested in research. Until we have dependable solutions of them, or at least critically tested hypotheses, we will not have a foundation for a real

science of education. Conventional test construction, fact-finding surveys, and curriculum analyses are, in contrast, secondary and in many instances trivial. The findings of such investigations may be helpful in planning educational practice but their contribution to a science of education is, in most cases, negligible, and in the few instances where a contribution is made, the results are so fragmentary that their value is doubtful. Hence, those who are interested in developing a science of education should endeavor to comprehend the fundamental problems and to contribute to their solution.

* * * * *

A concluding statement. Careful study of the preceding pages of this text should have revealed to the reader the position of its authors with respect to the relation of a science of education to the theory and practice of education. The fundamental problems of education were stated in Chapter II. The last of these problems was stated in terms of "what should be" and, in Chapter XII, the importance of philosophical thinking in dealing with this problem was stressed. It should be noted, however, that the solution of the other fundamental problems will yield generalizations whose application to educational practice involves consideration of "what should be." Research of the type of multiple factor analysis will result in conclusions respecting elemental human abilities and traits. These conclusions will be instrumental in the construction of better measuring instruments. More valid tests will aid in the attainment of more dependable experimental findings. Improved knowledge respecting human abilities and traits and the factors influencing them will aid in the formulation of more defensible educational objectives, and, hence, in the construction of more adequate curricula. There is a danger that research workers engaged in the attempted solution of aspects of the fundamental problems will lose sight of the intimate relation between a science of education and the practice of education. The philosophical implications of the fundamental research of the "pure science"

type cannot be left to philosophers alone. Too often educational philosophers are erratic in their thinking about generalizations derived from scientific research in education. Some of their criticisms have been well founded, but in numerous instances, these philosophers have formulated theories about education without careful consideration of the faults of the research conclusions used as data in their thinking. In many cases, research workers in education have become lost in their enthusiasm for techniques, particularly the statistical ones. "Statistical" significance has, for these individuals, become of greater importance than "practical" significance.

APPENDIX

STATISTICAL SYMBOLS

Symbols are used as a means of convenience in designating statistics and in expressing formulae. Since they are employed as instruments of communication, uniformity of usage is highly desirable. Unfortunately, authors of statistical texts, and other writers employing statistical symbols, exhibit many variations in usage.¹ This lack of uniformity increases the difficulty of reading reports of educational research. A number of symbols are listed in this appendix for the convenience of persons who desire to systematize their usage. The list has been compiled after considerable study of the matter and in most cases the symbols given are strongly supported by current practice. In a few cases a symbol not widely used is given because it appears to represent a simple practice. No attempt has been made to supply symbols for all statistics.

Unfortunately our symbolism has developed without much attention to general principles. A few principles, however, are rather generally observed by the more authoritative writers.

Designation of data: Test scores and other measures. A group of data, such as the scores made on a test, chronological ages, school marks, intelligence quotients, and the like, commonly thought of as representing values of a variable, is usually designated by the symbol, X . A second variable may be designated by Y , but most writers prefer to attach numerical subscripts to X to designate the several groups of data or variables. This practice has the advantage of being capable of extension to any number of variables. When only two sets of data are involved, the most common designations are X_1 and X_2 , but there are certain advantages in using X_0 to designate the variable that is considered criterion or dependent. The independent variables would be represented by $X_1, X_2, X_3 \dots X_n$.

When the raw data have been transformed so that they are expressed as deviations from their mean as the zero point, small letters are used

¹ West has reported the variation in usage in sixteen texts. He found different symbols used to represent the same thing and a number of statistics that were represented by two or more symbols.

West, P. V. "Need for Standardization of Symbols and Formulae in Educational Statistics," *Journal of Experimental Education*, 1: 216-22, March, 1933.

instead of capitals. If they are further transformed so that they are expressed in terms of their standard deviation (σ) as a unit, z is used as the symbol. When the data are expressed from an arbitrary zero point, such as an assumed mean, a prime ($'$) is attached to the symbol. A bar ($\bar{}$) above a symbol indicates a value estimated by means of a formula, usually a regression equation. An estimated value usually has the meaning of "most probable" value.

A with subscripts as needed, is used to designate a variable that is a component of X . When the variables are in terms of deviations, a is used for this purpose. Sometimes a, b, c, d , etc., are used instead of a with subscripts.

The number of measures (cases) in a group or population is commonly designated by N . Small n is used to represent the number of variables or sets of measures. Subscripts may be attached to N to designate different populations or groups. When only two populations are involved, n is sometimes used to designate the smaller one.

The results of calculations. The results of certain calculations such as those designed to obtain a central tendency or a measure of relationship are commonly designated by a single letter: M for mean, r for Pearson product-moment coefficient of correlation, σ for standard deviation, etc. There are a few exceptions of which the use of Md for median and PE for probable error are perhaps the most important.

The use of subscripts. It is usually desirable to connect a symbol designating the result of certain calculations with the group or groups of data from which it was obtained. This is accomplished by means of subscripts. For example, M_3 indicates the mean of the measures designated by X_3 ; r_{24} indicates the coefficient of correlation was obtained from the measures of variables, X_2 and X_4 or x_2 and x_4 . In writing regression equations, coefficients of partial correlation, and certain other statistics, the system of subscripts is rather elaborate, but when it is understood, the statistics are easy to write and read. The use of sub-subscripts should be avoided when feasible. Thus, we write r_{23} instead of $r_{X_2X_3}$, and D_{1-2} instead of $D_{M_1-M_2}$. When an additional subscript is required, it may be written as a prefix. Thus, we write ${}_2X_0$ rather than X_{0_2} and ${}_2r_{12}$ rather than r_{12_2} .

New symbols. A writer will facilitate the reading of his publications by employing the symbols that are sanctioned by general usage. It is unwise to follow a text or other source that does not represent good practices. When a writer encounters a need for new symbols, he should be guided by two general rules: (1) Avoid, if possible, the use of a letter or other symbol that has any considerable usage for another purpose. (2) Select a simple symbol. In general, use a single letter, with an appropriate subscript if necessary. When two letters are used, omit periods.

A LIST OF RECOMMENDED SYMBOLS

- AA* Achievement age, or synonymously accomplishment age or attainment age.
- AD* Average deviation, or mean deviation.
- AQ* Achievement quotient.
- AR* Achievement ratio.
- b_{01} Regression coefficient, involving r (Pearson) unless otherwise specified, of X_0 (dependent) on X_1 (independent) or x_0 on x_1 . $\left(b_{01} = r_{01} \frac{\sigma_0}{\sigma_1}\right)$. See page 324.
- $b_{01.23 \dots n}$ Partial regression coefficient of X_0 on X_1 , all others of the n independent variables constant. See page 325.
- $\beta_{01.23 \dots n}$ Beta regression coefficient. See page 325.
- C* Constant in a regression equation. Used as subscript designates control group.
- CA* Chronological age.
- CR* Critical ratio. $CR = \frac{D_{M_1 - M_2}}{PE_D}$. See *EC*.
- c* Correction, difference between assumed mean and exact mean. As a subscript before σ_1 and r_{12} indicates that a correction for coarse grouping has been employed.
- D* Difference. The quantities subtracted may be indicated by a subscript. For example, D_{10-90} designates the 10-90 percentile range.
- d* Deviation from the mean in terms of class or step intervals. The preferred symbol for this purpose is x . It should always be used when the deviation is in terms of scale units. A prime (') is attached to indicate a deviation from an assumed or arbitrary origin.
- d_{01} Coefficient of determination used in connection with path coefficients. See page 393.
- d_{012} A coefficient of determination measuring the joint effect of variables x_1 and x_2 on the variance of x_0 .
- E* Efficiency of prediction. For specific meanings as indicated by subscripts see pages 340 f.

e	Designates the base of the Naperian system of logarithms and equals 2.71828 . . . Also used to designate the error in a measure.
$e_{\hat{v}ar}$	Variable error of measurement.
e_{sys}	Systematic error.
EA	Educational age, expresses standing in a number of school subjects.
EC	Experimental coefficient proposed by McCall.
EC	$= \frac{D_{M_1 - M_2}}{2.78\sigma_D}$. See CR .
EQ	Educational quotient.
f	Frequency within an interval.
G	Gain.
IQ	Intelligence quotient.
i	Width of class or step interval in scale units.
K	A constant. (Not used in regression equations.)
k	Coefficient of alienation. $k = \sqrt{1 - r^2}$. When squared is the coefficient of non-determination. See r for subscripts and their meaning.
$k_{1.23 \dots n}$	Multiple alienation coefficient. When squared is the multiple coefficient of non-determination.
$k_{12.34 \dots n}$	Partial alienation coefficient.
M	Mean.
MA	Mental age.
Md	Median.
MdD	Median deviation.
MD	Mean deviation. This symbol is <i>not</i> recommended. Use AD instead.
Mo	Mode. Sometimes designated by Z .
m	Number of variables when n is used in same formula.
N	Total number of cases or observations.
n	Number of variables. Also number of alternative responses to an item on a multiple response test.

- O* Used as a subscript to designate a criterion or dependent variable.
- P* With subscripts from 1 to 100 designates a percentile point. For example, P_{10} designates the tenth percentile, the point on the scale of a frequency distribution below which 10 per cent of the measures fall and P_{90} designates the ninetieth percentile, the point on the scale of a frequency distribution below which 90 per cent of the measures fall.
- PE* Probable error, median deviation of the distribution when it is one of errors. $PE = .6745\sigma$ where σ represents the standard error. *PE* is sometimes used incorrectly for *MdD* where the distribution is not one of errors.
For the use of this symbol with subscripts to designate the probable error of particular quantities see σ (standard error).
- p* Probability of success, or of per cent of cases in a given category. $p = 1 - q$.
- p_{01} Symbol for path coefficient. See page 393.
- Q* Quartile deviation. Sometimes called semi-interquartile range. $Q = \frac{Q_3 - Q_1}{2}$
- Q_1 First (lower) quartile point. $Q_1 = P_{25}$.
- Q_3 Third (upper) quartile point. $Q_3 = P_{75}$.
- q* Probability of failure. $q = 1 - p$.
- $R_{1.234 \dots n}$ Multiple correlation coefficient. When squared it is the coefficient of multiple determination.
- R_{12} Pearson product-moment coefficient of correlation in a theoretical range of talent. Usually this range of talent is larger than that for r_{12} . *R* is used as a symbol for the coefficient of "rank correlation," but this statistic is seldom calculated. *R* is also used to designate the "number of right responses."
- r* Pearson product-moment coefficient of correlation. Subscripts are used to indicate the two sets of paired measures or variables whose correlation is being expressed. The symbols X_1, X_2 , etc., designating the

variables may be used as subscripts but usually only the subscripts of these symbols are attached to r . Usually no significance is attached to the order in which the two subscripts are written, but when it is desired to identify the dependent variable (Y -ordinate in correlation table) the first position may be given to the subscript designating this variable. A few of the more common cases of special subscripts are given below.

r_{II}	Coefficient of reliability, the subscripts designating two measures of the same thing. We also write r_{2II} . See pages 199 f.
$r_{\frac{1}{2} II}$	Coefficient of correlation between the two halves of a test. See page 201.
$r_{1\infty}$	Correlation between obtained and theoretical true measures. This symbol assumes that X_1 represents the obtained measures. If another symbol is used to designate these measures the subscript " 1 " would be changed accordingly. When X_1 designates test scores $r_{1\infty}$ is read "index of reliability."
$r_{\infty\omega}$	Coefficient of correlation corrected for attenuation. See page 151.
$r_{12\cdot34} \dots n$	Coefficient of partial correlation. See pages 377 f.
S	Sometimes used to designate summation. See Σ .
SD	Standard deviation, but this symbol is seldom used. See σ .
s	Used as subscript to indicate errors due to random sampling.
sys	Used as subscript to designate systematic error.
Sk	Skewness.
t_{1234}	Symbol for tetrad difference. $t_{1234} = r_{12}r_{34} - r_{13}r_{24}$. See page 402.
V	Coefficient of variability; a measure of relative dispersion. Often written $C.$ of V .
var	Used as subscript to designate variable error.
X and Y	Are used to designate raw or observed measures in two series of measures, or X_1 and X_2 may be used.

If there are several series of raw measures these may be designated $X_1, X_2, X_3, \dots X_n$. Small letters x, y, x_1, x_2, x_3 and so on represent measures expressed as deviations from the means of the corresponding raw measures. (See introductory statement.)

X_∞ Symbol for true measures corresponding to the raw or obtained measures designated by X . If there are two or more X 's such as X_0, X_1, X_2, \dots , the corresponding true measures may be designated by ${}_\infty X_0, {}_\infty X_1, {}_\infty X_2 \dots$ as a means of avoiding confusion.

\bar{X}_0 Symbol for predicted or estimated measure of variable X_0 . Similarly, $\bar{X}_1, \bar{X}_2, \bar{X}_3$, etc., represent predicted or estimated measures of variables X_1, X_2, X_3 , etc., where these are considered as *dependent* variables.

\bar{X}_∞ *Estimated* true score, the variable is X_1 , when it is not, the symbols ${}_\infty \bar{X}_0, {}_\infty \bar{X}_2$, etc., may be employed. See page 152.

z Standard measure, i.e., a measure expressed from the mean of the distribution as a zero point and in terms of the standard deviation (σ) as a unit.

$z_1 = \frac{X_1 - M_1}{\sigma_1}$. The symbol z is also used as an ordinate of the normal probability curve having unit area and unit standard deviation.

η Eta, the ratio of correlation; measure of curvilinear correlation. See page 98.

Σ Summation, the sum of. Occasionally S is used for this purpose but except where special designation is necessary, Σ is preferable as a symbol. Σ is sometimes used to indicate standard deviation of theoretical or large range.

\sum_{1}^N The sum of the measures for individuals, 1 to N inclusive

σ Sigma, the standard deviations of a distribution.

$\sigma = \sqrt{\frac{\sum x^2}{N}}$. See pages 75 f. When the distribution is

one of errors, σ is called the standard error. See pages 104 f. and 133. The particular distribution is indicated by subscripts. A few special cases are given.

σ_M Standard error of a mean, usually interpreted as the standard error due to random sampling. If the data are fallible, the effect of variable errors of measurement is also included. See pages 156–57. When there is any possibility for confusion, sub-subscripts should be used as in the three following cases.

$\sigma_{M_{s+e}}$ Standard error of a mean due to random sampling and to variable errors of measurement.

σ_{M_e} Standard error of a mean due to variable errors of measurement.

σ_{M_s} Standard error of a mean due to random sampling.

σ_{Md} Standard error of a median.

σ_r Standard error of a coefficient of correlation.

σ_D Standard error of a difference. When desired the subscript D may be replaced by the difference it represents. Hence $\sigma_{M_1 - M_2}$ would indicate the standard error of the difference $M_1 - M_2$. When necessary to avoid confusion $r_{M_1 M_2} \neq 0$ or $r_{M_1 M_2} = 0$ may be added to the subscript to make clear the formula used. See page 105.

σ_∞ True standard deviation or standard deviation of true measures. $\sigma_\infty = \sigma_1 \sqrt{r_{1I}}$. When desirable to indicate the group of data or variable the symbol ∞ may be written as a pre-subscript.

$\sigma_{1.2}$ Standard error of estimate—standard deviation of the differences between X_1 and the estimates of $X_1(\bar{X}_1)$ made from X_2 by simple regression equation, $\bar{X}_1 = r_{12} \frac{\sigma_1}{\sigma_2} X_2 + C$. We might write $\sigma_{X_1 - \bar{X}_1}$ but this would not be a convenient symbolism. Hence we use $\sigma_{1.2}$ in which $_2$ indicates the basis of estimate and $_1$ the basis of comparison. $\sigma_{1.2} = \sigma_1 \sqrt{1 - r_{12}^2}$. In case the variables are X_0 and X_1 the standard error of estimate would be expressed as $\sigma_{0.1}$. See pages 334 f.

- $\sigma_{2.1}$ Standard error of estimate of \bar{X}_2 , the predictions being made from X_1 by means of regression equation.
 $\sigma_{2.1} = \sigma_2 \sqrt{1 - r_{12}^2}$.
- $\sigma_{\infty.1}$ Standard error of estimate where a regressed, or estimated true score, \bar{X}_{∞} is taken as evidence of the true score X_{∞} . ($X_1, \bar{X}_{\infty}, X_{\infty}$ represent the same function.) $\sigma_{\infty.1} = \sigma_1 \sqrt{r_{1I}^2 - r_{1I}^2}$. The symbol is also used where \bar{X}_0 is taken as evidence of $_{\infty}X_0$, and the formula for the standard error of estimate is $\sigma_0 \sqrt{r_{00}^2 - r_{01}^2}$. See page 336.
- $\sigma_{1.\infty}$ Standard error of measurement where X_1 is taken as evidence of X_{∞} . Literally the standard error of measurement is the standard deviation of the difference $X_1 - X_{\infty}$ (variable error of measurement).
 $\sigma_{1.\infty} = \sigma_1 \sqrt{1 - r_{1I}^2}$. See pages 132-33.
- $\sigma_{0.1234 \dots n}$ Standard error of estimate of \bar{X}_0 computed from regression equation involving independent variables $X_1, X_2, X_3, \dots n$. Same symbol is used when the variables are expressed from their respective means as zero points.
- $\sigma_{\infty.1234 \dots n}$ Standard error of estimate where \bar{X}_0 is taken as evidence of true criterion measure $_{\infty}X_0$, multiple regression equation being used with $X_1, X_2, X_3 \dots X_n$ as independent variables.
- ∞ Infinite symbol, as a subscript indicates a true measure of a variable, i.e., the mean of an infinite number of measurements.
- ω Omega, as a subscript, a true measure of a second variable.

AUTHOR INDEX

- Aamodt, Geneva P., and Torgerson, T. L., 349
 Abelson, H. H., 11
 Abelson, Paul, 168
 Ackerson, Luton, 94
 Ackerson, Luton; Rich, G. M.; and Jackson, J. D., 201
 Adams, H. F., 407
 Advisory Committee on College Testing, 241, 242
 Alexander, Carter, 12, 25, 31, 239, 440, 442, 448
 Alexander, Carter, and Manske, A. J., 442
 Allen, C. B., 239
 Allen, Jerome, 453
 Almack, J. C., 12
 Anastaci, Anne, 198
 Anderson, C. A., 53
 Anderson, E. M., 259
 Anderson, H. A., and Traxler, A. E., 178
 Anderson, H. R., and Lindquist, E. F., 172, 184
 Anderson, L. D., and Toops, H. A., 95
 Anderson, P. L., and Yankey, J. V., 452
 Andrus, Ruth, 259
 Anibel, F. G., 295
 Ashbaugh, E. J., 131
 Asher, Ollie, and Monroe, W. S., 441
 Ayer, F. C., 15
 Ayres, L. P., 94, 210, 222, 426, 455, 456
 Baehne, G. W., 228
 Bagley, W. C., 426
 Bagley, W. C., and Rugg, H. O., 426
 Bailor, E. M., 342, 343
 Bain, A., 453
 Baird, D. O., 259
 Bakst, A., 331
 Baldwin, B. T., 407
 Bamesberger, V. C., 426
 Banker, H. J., 222
 Barlow, 64
 Barlow, M. C., and Peterson, Joseph, 189
 Barnett, N. E., and Potthoff, E. F., 190
 Barr, A. S., 25, 259, 306, 370
 Barr, A. S., and Douglas, Lois, 353
 Barr, A. S., and Gifford, C. W., 259
 Barr, A. S., and Park, J. S., 314
 Barr, A. S., and Rudisill, Mabel, 12
 Barr, A. S., et al., 174
 Barthelmess, H. M., 185
 Bartlett, L. W., 260
 Baten, W. D., 379
 Bawden, W. T., 41
 Beatley, Bancroft, 111
 Bender, J. F., 260
 Bennett, C. C., and Gates, A. I., 283
 Bennett, H. E., 260
 Bergman, W. G., and Vreeland, Wendell, 315
 Bernheim, Ernst, 160, 164
 Betts, E. A., and Greene, H. A., 47
 Binet, A., and Simon, T., 183, 210, 456
 Bixler, G. K., 427
 Bixler, H. H., 12, 36
 Bjarnason, Loftor, 50
 Blan, L. B., 455
 Blankenship, A. S., 260
 Blume, C. E., 50
 Boardman, C. W., 354
 Bobbitt, Franklin, 426, 427
 Bobbitt, Franklin, et al., 35, 427
 Bobbitt, Sarah, 427

- Bode, B. H., 415, 421, 423, 427, 433, 471
Bogardus, E. S., 36
Book, W. F., and Harter, R. S., 261
Bovard, J. F., and Cozens, F. W., 210
Bowden, A. O., 427
Bowley, H. L., 149
Breed, F. S., 294
Bregman, E. O., and Thorndike, E. L., 268
Breslich, E. R., 294
Bridges, K. M. B., 50
Briggs, T. H., 25, 428
Briggs, T. H., et al., 26
Brigham, C. C., 185
Brinkemeier, I. H., and Keys, Noel, 145
Brinkemeier, I. H., and Ruch, G. M., 145
Brintle, S. L., and Segel, David, 364
Brinton, W. C., 118, 239
Brooks, E. C., 288
Brooks, F. D., 296
Brown, A. E., 315
Brown, S. W., 168
Brown, William, 201
Brown, William, and Thomson, G. H., 111, 152, 210
Brownell, W. A., 202, 301, 312, 370, 451, 466
Brueckner, L. J., 52
Bruner, H. B., and Stratemeyer, F. B., 433
Buckingham, B. R., 210, 414, 459
Buckner, M. A., 261
Burgess, May Ayres, 188
Burgess, W. R., 222
Burk, Frederic, 40
Burks, B. S., 395, 407
Burks, J. D., and Stone, C. R., 295
Burns, R. L., 221
Buros, O. K., 211
Burt, C., 389
Buswell, G. T., 22, 438
Buswell, G. T., and John, Lenore, 316
Buswell, G. T., and Judd, C. H., 3, 8, 9, 319, 438, 451
Butterfield, E. W., 40
Cady, V. M., 197, 203
Cairns, George J., 17, 404, 407
Caldwell, O. W., and Finley, C. W., 35, 429
Caldwell, O. W., and Lundeen, G. E., 264
Caliver, Ambrose, 261
Camp, B. H., 97, 106, 118
Carpenter, H. S., and Elder, Vera, 262
Carroll, H. A., and Hollingworth, L. S., 141
Cason, Hulsey, 303
Caswell, H. L., 214, 216
Cattell, J. McK., 143, 454
Cattell, J. McK., and Farrand, Livingston, 454
Cattell, Psyche, 141, 316
Chaddock, R. E., 104, 118, 224, 232, 239, 250, 323, 330
Chadsey, C. E., et al., 2, 261
Chamberlain, L. M., and Crawford, A. B., 238, 239
Chambers, O. R., 278
Chant, S. N. F., 404
Chapin, F. S., 54
Chapman, J. C., and Eby, H. L., 261
Chapman, J. C., and Sims, V. M., 6, 54
Charters, W. W., 35, 316, 428, 466
Charters, W. W., and Waples, Douglas, 428
Charters, W. W., and Whitley, I. B., 428, 434
Cheshire, Leone; Saffir, Milton; and Thurstone, L. L., 96
Childs, H. G., and Terman, L. M., 212
Christofferson, H. C., 261
Clark, H. F., 222, 223
Clark, J. A., and Monroe, W. S., 36, 431
Clark, J. R., and Vincent, E. L., 131

- Clark, Mildred, and Worcester, D. A., 316
 Clark, W. W., 52
 Clark, W. W., and Williams, J. H., 54
 Clayton, Blythe, and Holzinger, K. J., 201
 Clugston, H. A., and Davis, R. A., 414
 Cochran, R. E., and Weidemann, C. C., 180, 303
 Cocking, W. D., 428
 Cogan, L. C.; Conklin, A. M.; and Hollingworth, H. L., 143
 Cole, P. R., 168
 Collings, Ellsworth, 285; 316, 431
 Collins, J. E., 408
 Commins, W. D., 280
 Committee of the Manchester Statistical Society, 40
 Conklin, A. M.; Cogan, L. C.; and Hollingworth, H. L., 143
 Conrad, H. S., and Martin, G. B., 345
 Cook, W. W., 184
 Cook, W. W., and Lindquist, E. F., 184, 185, 210
 Corey, S. M., 190, 281, 302, 354, 387, 451
 Cornell, E. L.; Coxe, W. W.; and Orleans, J. S., 53
 Coryell, N. G., 300
 Counts, G. S., 54, 261, 429
 Courtis, S. A., 275, 276, 295, 456, 458, 463, 467, 468, 471
 Courtis, S. A., and Thorndike, E. L., 145
 Courtis, S. A., et al., 210
 Coxe, W. W.; Cornell, E. L.; and Orleans, J. S., 53
 Cozens, F. W., and Bovard, J. F., 210
 Cozens, F. W., and Douglass, H. R., 201, 204, 210
 Crabbs, L. M., 354
 Crawford, A. B., and Chamberlain, L. M., 238, 239
 Crawford, C. C., 12
 Crelle, A. L., 64
 Croce, Benedetto, 160
 Crooks, A. D., 451
 Croon, C. W., and Rulon, P. J., 298
 Crosby, Amy, and Hall, Irene, 279
 Cubberley, E. P., 455
 Cunningham, E. M., and Olson, W. C., 59
 Cureton, E. E., 95, 405
 Cureton, E. E., and Dunlap, J. W., 379, 382, 383, 384, 393
 Cureton, E. E.; Dunlap, J. W.; and Pratt, H. G., 275
 Curtis, F. D., 55, 438
 Daly, J. F., and Furfey, P. H., 87, 101
 Daniel, R. P., 200
 Davis, E. C., 216
 Davis, R. A., 26
 Davis, R. A., and Clugston, H. A., 414
 Davis, R. A., and Franke, P. R., 461
 Dearborn, N. H., 168
 Dearborn, W. F., 211, 455
 Dearborn, W. F., and Smith, W. C., 132
 Dech, A. O., 426
 Demiashevich, M. J., 453
 Denworth, K. M., 286, 408
 Derring, C. E., 442
 Dewey, John, 429, 433
 Dolch, E. W., 35, 59
 Donnelly, H. I., 197
 Douglas, Lois, and Barr, A. S., 353
 Douglass, H. R., 4, 26, 41, 174, 222, 254, 291, 295, 311, 317, 342, 351, 352, 419, 469
 Douglass, H. R., and Cozens, F. W., 201, 204, 210
 Douglass, H. R., and Huffaker, C. L., 156, 310
 Douglass, H. R., and Spencer, P. L., 190
 Dow, E. W., 163
 Droba, D. D., 197
 Dulebohn, I. H., 427

- Dunlap, J. W., 200, 236
 Dunlap, J. W., and Cureton, E. E.,
 379, 382, 383, 384, 393
 Dunlap, J. W., and Kurtz, A. K.,
 64, 74, 78, 88, 100, 118, 151
 Dunlap, J. W., and McNamara,
 W. J., 184, 236
 Dunlap, J. W.; Pratt, H. G.; and
 Cureton, E. E., 275
 Dvorak, August, 99
 Dyer, C. A., 427
 Dynes, J. J., 317

 Eason, J. L., 429
 Eaton, M. T., 318
 Eby, H. L., and Chapman, J. C., 261
 Edgerton, H. A., 356
 Edgerton, H. A., and Toops, H. A.,
 202, 209
 Eells, W. C., 75
 Elder, Vera, and Carpenter, H. S.,
 262
 Elliott, E. C., and Starch, Daniel,
 321, 455
 Elsbree, W. S., 54, 262
 Elsbree, W. S.; Strayer, G. D.; and
 Engelhardt, N. L., 54
 Engelhardt, Fred, 238
 Engelhardt, N. L., and Strayer,
 G. D., 54
 Engelhardt, N. L.; Reeves, C. E.;
 and Womrath, G. F., 54
 Engelhardt, N. L.; Strayer, G. D.;
 and Elsbree, W. S., 54
 Engelhart, M. D., 295, 451
 Engelhart, M. D., and Monroe,
 W. S., 272, 309, 449, 452
 Engelhart, M. D.; Monroe, W. S.;
 Odell, C. W.; Herriott, M. E.;
 and Hull, M. R., 12, 167, 183
 Eurich, A. C., and Johnson, D. A.,
 60
 Eurich, A. C., and Kinney, L. B.,
 208
 Ezekiel, Mordecai, 109, 118, 309,
 326, 347, 356, 360, 361, 371, 383
 Ezekiel, Mordecai, and Tolley,
 H. R., 331

 Fallon, J. F., 41
 Farnsworth, P. R., 201
 Farrand, Livingston, and Cattell,
 J. McK., 454
 Feldstein, M. J., 96
 Finch, F. H., and Stokes, C. N.,
 248
 Finch, J. H.; Lentz, T. F.; and
 Hirshstein, Bertha, 184
 Findlay, J. J., 453
 Finley, C. W., and Caldwell, O. W.,
 35, 429
 Fisher, Irving, 119, 224
 Fisher, R. A., 109, 119, 252, 382
 Fitch, H. N., 262
 Fitzpatrick, E. A., 168
 Flaccus, Quintus H. (Pseudonym),
 40
 Fleming, C. W., 279
 Fling, F. M., 160, 164
 Flowers, I. V., 51, 262
 Foran, T. G., 203, 208
 Foster, J. C., 262
 Fowlkes, J. G., 34, 257
 Franke, P. R., and Davis, R. A.,
 461
 Franzen, R. H., 383
 Franzen, R. H., and Knight, F. B.,
 34, 53
 Freeman, E. A., 164
 Freeman, F. N., 15, 211, 415, 464,
 471
 Freeman, F. N., and Holzinger,
 K. J., 389, 390
 Freeman, F. N., and Wood, B. D.,
 321
 Freyd, Max, 52
 Fryer, Douglas, 211, 279, 451
 Fuller, L. R., 429
 Furfey, P. H., and Daly, J. F., 87,
 101

 Galton, Francis, 366, 454
 Garretson, O. K., 197
 Garrett, H. E., 80, 107, 119, 225,
 249, 331, 390, 403
 Gatchel, D. F., 287
 Gates, A. I., 129, 188, 278

- Gates, A. I., and Bennett, C. C., 283
George, H. B., 164
Gesell, Arnold, 49
Gevorkiantz, S. R., and Mudgett, B. D., 105
Gifford, C. W., and Barr, A. S., 259
Gilliland, A. R., and Misbach, L. E., 302
Goddard, H. H., 183
Good, C. V., 12, 26
Good, H. G., 161, 164, 169
Goodenough, F. L., and Terman, L. M., 55
Gordon, H. C., 362
Grant, A., and Remmers, H. H., 34, 266
Gray, C. T., 3
Gray, J. S., 10
Gray, W. S., 50, 312, 438, 466
Greene, H. A., and Betts, E. A., 47
Greenleaf, W. J., and Windes, E. E., 443
Griffin, Harold D., 65, 329, 331, 332, 397
Grinstead, W. J., 311
Grover, C. C., 362
Gulliksen, H.; Wilson, W. R.; and Welsh, G., 184
Gwynn, Aubrey, 169

Haefner, Ralph, 298
Haggerty, M. E., 355
Haggerty, M. E.; Olson, W. C.; and Wickman, E. K., 53
Hall, G. S., 454
Hall, Irene, and Crosby, Amy, 279
Hall, J. J., 54
Hall-Quest, A. L., 34
Hamilton, T. T.; Monroe, W. S.; and Smith, V. T., 441, 442
Handy, Urvan, and Lentz, T. F., 204
Hankinson, Frank, 40
Hansen, A. O., 169
Harap, Henry, 430
Harap, Henry, and Persing, E. C., 430

Hardy, M. C., and Hoefer, Carolyn, 279
Harter, R. S., and Book, W. F., 261
Hartmann, G. W., 466
Hartshorne, Hugh, and May, M. A., 211
Haught, F. B., 53
Hayes, H., and Sturtevant, S. M., 36
Heck, A. O., 257
Hedman, H. B., and Line, W., 402
Heilman, J. D., 5, 8, 9, 55, 83, 219, 279, 395
Helseth, Inga Olla, 47
Henderson, E. N., 167
Hendrickson, Gordon, 171, 471
Henmon, V. A. C., 26
Henry, N. B., 262
Herring, J. P., 205
Herriott, M. E., 52, 278, 408
Herriott, M. E., and Monroe, W. S., 431
Herriott, M. E.; Monroe, W. S.; Odell, C. W.; Engelhart, M. D.; and Hull, M. R., 12, 167, 183
Herron, J. S., and Sexton, E. K., 4, 8, 285
Hewitt, Alden, 263
Hildreth, G. H., 211
Hillegas, M. B., 211, 456
Hilliard, G. H., 451
Hindman, D. A.; Monroe, W. S.; and Lundin, R. S., 431
Hirshstein, Bertha; Lentz, T. F.; and Finch, J. H., 184
Hockett, J. A., 6, 8, 9, 423, 430
Hoefer, Carolyn, and Hardy, M. C., 279
Hoffman, G. J., 143
Hoke, K. J., and Wilson, G. M., 213
Holley, C. E., 54
Hollingworth, H. L., 53
Hollingworth, H. L.; Conklin, A. M.; and Cogan, L. C., 143
Hollingworth, L. S., and Carroll, H. A., 141
Holzinger, K. J., 64, 79, 80, 84, 88, 92, 97, 99, 100, 102, 111, 119, 128,

- 130, 150, 194, 201, 205, 209, 236,
241, 275, 324, 326, 338, 342, 379,
388, 404, 408
- Holzinger, K. J., and Clayton,
Blythe, 201
- Holzinger, K. J., and Freeman,
F. N., 389, 390
- Holzinger, K. J., and Swineford,
Frances, 401
- Hopkins, L. T., 430
- Horn, Ernest, 47, 49, 430
- Horst, Paul, 84, 85, 331
- Horton, R. E., 318
- Hotelling, Harold, 401
- Hudelson, Earl, 287, 318, 451
- Huffaker, C. L., 95, 205
- Huffaker, C. L., and Douglass,
H. R., 156, 310
- Hull, C. L., 83, 96, 211, 224, 329,
357, 361, 379
- Hull, M. R.; Monroe, W. S.; Odell,
C. W.; Herriott, M. E.; and
Engelhart, M. D., 12, 167, 183
- Hullfish, H. G., 415
- Hunt, Thelma, 353
- Hurd, A. W., 287
- Hurlock, E. B., 143, 318
- Inman, J. H., 263
- Irion, T. W. H., 56, 248
- Jackson, G. L., 169
- Jackson, J. D.; Ruch, G. M.; and
Ackerson, Luton, 201
- James, William, 454
- John, Lenore, and Buswell, G. T.,
316
- Johnson, Allen, 164
- Johnson, D. A., and Eurich, A. C.,
60
- Johnston, J. B., et al., 242
- Jones, D. C., 119
- Jordan, A. M., 208
- Jordan, R. C., 203
- Judd, C. H., 26, 419, 457, 469
- Judd, C. H., and Buswell, G. T.,
3, 8, 9, 319, 438, 451
- Justice, W. A., 92
- Kaplan, E., and Line, W., 403
- Kaplan, E.; Line, W.; and Rogers,
K. H., 409
- Karsten, K. J., 119, 239
- Kaulfers, W. V., 341, 351, 362
- Kefauver, G. N., 85, 141
- Kelley, T. L., 12, 25, 80, 92, 97, 100,
101, 104, 110, 111, 119, 150, 151,
152, 156, 190, 194, 200, 201, 202,
204, 211, 224, 236, 325, 332, 335,
363, 366, 374, 376, 401, 402, 403,
405, 408, 409, 415, 423
- Kelley, T. L., and Krey, A. C., 176
- Kelley, T. L.; Ruch, G. M.; and
Terman, L. M., 135
- Kellogg, C. E., and Spence, K. W.,
150
- Kelly, E. L., 393
- Kelly, E. L., and Whitney, F. L.,
263
- Kelly, E. L.; Remmers, H. H.; and
Shock, N. W., 201
- Kelly, F. J., 26, 263, 455
- Kemp, W. W., 169
- Kennedy, L. R., 173
- Keyes, C. H., 455
- Keys, Noel, and Brinkemeier,
I. H., 145
- Kilpatrick, W. H., 415, 431, 433
- Kinney, L. B., and Eurich, A. C.,
208
- Klapper, Paul, 26
- Klein, A. J., et al., 263
- Knight, E. W., 169
- Knight, F. B., 53, 142, 282, 285
- Knight, F. B., and Franzen, R. H.,
34, 53
- Knowlton, D. C., and Tilton, J. W.,
319
- Knudsen, C. W., 47, 50, 451
- Koos, L. V., 41
- Kornhauser, A. W., 54, 287
- Krey, A. C., and Kelley, T. L., 176
- Kuhlman, F., 183, 456
- Kulp, D. H., 26
- Kurtz, A. K., and Dunlap, J. W.,
64, 74, 78, 88, 100, 118, 151
- Kwalwasser, Jacob, 212

- Lamprecht, S. P., 366
Langlois, C. V., and Seignobos, C.,
160, 164
Lanier, L. H., 201
Larson, E. L., 244
Larson, S. C., 362
Lauer, A. R., 92, 102
Layton, E. T., 173
Lee, E. A., 26
Lee, J. M., and Symonds, P. M.,
208, 451
Lehman, H. C., 427
Lehman, H. C., and Stoke, S. M., 42
Lehman, H. C., and Witty, P. A.,
437
Lentz, T. F., and Handy, Urvan,
204
Lentz, T. F.; Hirshstein, Bertha;
and Finch, J. H., 184
Leonard, J. P., 319, 451
Lepley, Ray, 414, 419
Lien, Agnes, and Melby, E. O., 296
Limp, C. E., 363
Lincoln, E. A., 107, 208
Lincoln, E. A., and Shields, F. J.,
197
Lindquist, E. F., 248, 251, 294, 309,
315
Lindquist, E. F., and Anderson,
H. R., 172, 184
Lindquist, E. F., and Cook, W. W.,
184, 185, 210
Lindsay, E. E., 45
Line, W., and Hedman, H. B., 402
Line, W., and Kaplan, E., 403
Line, W.; Rogers, K. H.; and Kap-
lan, E., 409
Lively, B. A., and Pressey, S. L.,
34, 57
Long, J. A., 185
Longshore, W. T., et al., 264
Lorenzen, C. H., 427
Lundeen, G. E., and Caldwell,
O. W., 264
Lundin, R. S.; Hindman, D. A.;
and Monroe, W. S., 431
Lyman, R. L., 438, 452
Lyon, V. E., 252
McCall, W. A., 83, 187, 194, 212,
308, 389
McGaughy, J. R., 249, 264
McIntosh, H. W., and Schrammel,
H. E., 264
McNamara, W. J., and Dunlap,
J. W., 184, 236
Maddox, C. R., 50
Maddox, W. A., 169
Mahan, T. J., 431
Maller, J. B., 320
Mallory, J. N., 279
Manske, A. J., and Alexander,
Carter, 442
Marshall, R. L., 164
Martin, C. W., 26
Martin, G. B., and Conrad, H. S.,
345
Martz, H. B., and Peters, C. C., 302
Mathews, C. O., 42, 143
Maxwell, C. R., 34
Maxwell, C. R., et al., 264
Maxwell, W. H., 454
May, M. A., 377
May, M. A., and Hartshorne, Hugh,
211
Mead, A. R., 26
Melby, E. O., and Lien, Agnes, 296
Melchior, W. T., 265
Mendenhall, R. M., and Warren,
R., 65, 96
Meriam, J. L., 353, 431
Meyer, George, 147
Meyer, Max, 455
Mill, J. S., 366, 369, 370
Miller, L. W., 363
Miller, M. C., and Witmer, E. M.,
441
Miller, W. S., 85
Mills, F. C., 119
Miner, J. R., 64, 379
Misbach, L. E., and Gilliland, A. R.,
302
Monroe, Paul, 455
Monroe, W. S., 21, 26, 48, 134, 135,
136, 137, 138, 140, 177, 187, 190,
192, 194, 195, 212, 224, 265, 288,
297, 372, 441, 468

- Monroe, W. S., and Asher, Ollie, 441
 Monroe, W. S., and Clark, J. A., 36, 431
 Monroe, W. S., and Engelhart, M. D., 272, 309, 449, 452
 Monroe, W. S., and Herriott, M. E., 431
 Monroe, W. S., and Shores, Louis, 168, 441
 Monroe, W. S., and Souders, L. B., 135, 302
 Monroe, W. S., and Stuit, D. B., 111
 Monroe, W. S.; Hamilton, T. T.; and Smith, V. T., 441, 442
 Monroe, W. S.; Hindman, D. A.; and Lundin, R. S., 431
 Monroe, W. S.; Odell, C. W.; Herriott, M. E.; Engelhart, M. D.; and Hull, M. R., 12, 167, 183
 Moore, E. S., 54
 More, G. V. D., 363
 Morphet, E. L., 221
 Morris, E. H., 281, 353
 Morrison, H. C., 50, 291
 Mort, P. R., 26, 221
 Moul, M., and Pearson, Karl, 403
 Mudgett, B. D., and Gevorkiantz, S. R., 105
 Muthersbaugh, G. C., 432
 Nanninga, S. P., 409
 National Education Association, Research Division, 452
 Neivens, L. F., and Weidemann, C. C., 303
 Nelson, M. G., 265
 Nelson, M. J., 320
 Nietzsche, J. A., 427
 Noble, S. G., 169
 Noffsinger, F. R., and Scates, D. E., 190
 Norton, John K., 26, 41, 222
 Odell, C. W., 98, 112, 115, 119, 139, 189, 190, 193, 212, 241, 265, 286, 351, 352
 Odell, C. W.; Monroe, W. S.; Herriott, M. E.; Engelhart, M. D.; and Hull, M. R., 12, 167, 183
 Odell, W. R., 220
 Ohmann, O. A., 279
 Ojemann, R. H., 265
 Olander, H. T., 295, 320
 Olson, W. C., 53
 Olson, W. C., and Cunningham, E. M., 59
 Olson, W. C.; Haggerty, M. E.; and Wickman, E. K., 53
 Orleans, J. S., 95
 Orleans, J. S.; Cornell, E. L.; and Coxe, W. W., 53
 Osburn, W. J., 302, 428
 Otis, A. S., 92, 119, 183, 195, 212, 241
 Painter, W. I., and Patty, W. W., 58, 222
 Palmer, P. L., 427
 Park, J. S., and Barr, A. S., 314
 Parr, R. M., and Spencer, M. A., 284
 Paterson, D. G., 184, 302
 Paterson, D. G., et al., 364
 Patty, W. W., and Painter, W. I., 58, 222
 Payne, E. G., 27
 Payne, J., 453
 Payne, W. H., 453
 Pearson, Karl, 64, 79, 86, 111
 Pearson, Karl, and Moul, A., 403
 Pearson, Karl, et al., 153
 Peatman, J. G., 190
 Pechstein, L. A., 27
 Peik, W. E., 432
 Perry, H. E., 41
 Persing, E. C., and Harap, Henry, 430
 Peters, C. C., 415, 432, 433
 Peters, C. C., and Martz, H. B., 302
 Peters, C. C., and Van Voorhis, W. R., 108, 309
 Peters, C. C., and Wykes, E. C., 331
 Peters, J., 64
 Peterson, Joseph, 189, 212
 Peterson, Joseph, and Barlow, M. C., 189
 Phillips, D. E., 453

- Phillips, F. M., 31, 222
Pintner, Rudolph, 132, 212, 456
Pintner, Rudolph, and Thomson, G. H., 219, 388
Pittman, M. S., 285
Potthoff, E. F., and Barnett, N. E., 190
Powers, S. R., 452
Pratt, H. G.; Dunlap, J. W.; and Cureton, E. E., 275
Pressey, S. L., 279
Pressey, S. L., and Lively, B. A., 34, 57
Price, R. R., 266
Pu, A. S. T., and Trow, W. C., 143
Puckett, R. C., 49

Rabenort, W. L., 169
Rankin, P. T., 311
Reagan, G. W., 433
Reavis, G. H., 389
Reavis, W. C., 288, 389
Reckless, W. C., and Smith, Mapheus, 48
Reeder, E. H., 297, 321
Reeves, C. E.; Engelhardt, N. L.; and Womrath, G. F., 54
Reeves, F. W., and Russell, J. D., 222
Reigart, J. F., 169
Reisner, E. H., 415
Reitz, Wilhelm, 252
Remmers, H. H., 143, 201, 253
Remmers, H. H., and Grant, A., 34, 266
Remmers, H. H.; Shock, N. W.; and Kelly, E. L., 201
Research Division, National Education Association, 452
Rice, J. M., 271, 369, 453, 456, 458
Richardson, M. W., 184
Rietz, H. L., 120
Rietz, H. L., et al., 80, 120
Robbins, C. L., 169
Robinson, Eleanor, and Thorndike, E. L., 268
Rock, Jr., R. T., 452
Rogers, D. C., 84
Rogers, K. H., 403
Rogers, K. H.; Line, W.; and Kaplan, E., 409
Roller, Jr., R. D., and Stalnaker, E. M., 279
Ross, C. C., 364
Royce, J., 453
Ruch, G. M., 192, 212
Ruch, G. M., and Brinkemeier, I. H., 145
Ruch, G. M., and Stoddard, G. D., 193, 195, 212
Ruch, G. M.; Ackerson, Luton; and Jackson, J. D., 201
Ruch, G. M.; Kelley, T. L.; and Terman, L. M., 135
Ruckmick, C. A., 41
Rudisill, Mabel, and Barr, A. S., 12
Rufi, John, 39, 266
Rugg, H. O., 33, 52, 114, 231, 443, 454, 458
Rugg, H. O., and Bagley, W. C., 426
Rugg, H. O., et al., 433
Rulon, P. J., 321
Rulon, P. J., and Croon, C. W., 298
Rusk, R. R., 12
Russell, J. D., and Reeves, F. W., 222
Russell, O. R., 404

Saffir, Milton; Chesire, Leone; and Thurstone, L. L., 96
Savage, H. J., et al., 266
Scarf, R. C., 427
Scates, D. E., 10
Scates, D. E., and Noffsinger, F. R., 190
Schrammel, H. E., and McIntosh, H. W., 264
Schwegler, R. A., and Winn, Edith, 266
Scripture, E. W., 453
Segel, David, 65, 329, 333
Segel, David, and Brintle, S. L., 364
Seignobos, C., 164
Seignobos, C., and Langlois, C. V., 160, 164
Selke, Erich, 267

- Sexton, E. K., and Herron, J. S., 4, 8, 285
Seybolt, R. F., 165
Sharman, J. R., 39
Shen, Eugene, 53, 143, 201
Sheppard, W. F., 153
Shields, F. J., and Lincoln, E. A., 197
Shock, N. W.; Remmers, H. H.; and Kelly, E. L., 201
Shores, Louis, and Monroe, W. S., 168, 441
Simon, T., and Binet, A., 183, 210, 456
Simpson, A. D., 27
Sims, V. M., 302
Sims, V. M., and Chapman, J. C., 6, 54
Slocombe, C. S., 409
Smillie, W. G., and Spencer, C. R., 369
Smith, B. B., 361, 383
Smith, H. L., 267
Smith, J. G., 239
Smith, Mapheus, and Reckless, W. C., 48
Smith, Max, 187
Smith, V. T.; Monroe, W. S.; and Hamilton, T. T., 441, 442
Smith, W. C., and Dearborn, W. F., 132
Snedecor, G. W., 252, 409
Somers, T. T., 353
Souders, L. B., and Monroe, W. S., 135, 302
Spaulding, F. E., 34
Spearman, C., 151, 200, 201, 383, 402, 403, 404, 410
Spence, K. W., and Kellogg, C. E., 150
Spencer, C. R., and Smillie, W. G., 369
Spencer, M. A., and Parr, R. M., 284
Spencer, P. L., and Douglass, H. R., 190
Stalnaker, E. M., and Roller, Jr., R. D., 279
Stalnaker, J. M., and Stalnaker, R. C., 302
Starch, Daniel, and Elliott, E. C., 321, 455
Stenquist, J. L., 27
Stevens, Romiett, 47
Stoddard, G. D., and Ruch, G. M., 193, 195, 212
Stoke, S. M., and Lehman, H. C., 42
Stokes, C. N., and Finch, F. H., 248
Stone, C. R., and Burks, J. D., 295
Stone, C. W., 456
Strang, Ruth, 27
Strang, Ruth, and Sturtevant, S. M., 267
Stratemeyer, F. B., and Bruner, H. B., 433
Strayer, G. D., 27, 53, 455, 469
Strayer, G. D., and Engelhardt, N. L., 54
Strayer, G. D.; Engelhardt, N. L.; and Elsbree, W. S., 54
Strong, E. K., 434
Stuart, Hugh, 267
Stuit, D. B., 381
Stuit, D. B., and Monroe, W. S., 111
Sturtevant, S. M., and Hayes, H., 36
Sturtevant, S. M., and Strang, Ruth, 267
Swineford, Frances, and Holzinger, K. J., 401
Symonds, P. M., 15, 20, 27, 52, 53, 88, 130, 143, 175, 181, 184, 195, 196, 204, 212, 237, 278, 410, 417, 452, 468
Symonds, P. M., and Lee, J. M., 208, 451
Taylor, H., 354
Taylor, H. C., 170
Teggart, F. J., 160
Terman, L. M., 183, 212, 456, 466
Terman, L. M., and Childs, H. G., 212

- Terman, L. M., and Goodenough, F. L., 55
Terman, L. M.; Kelley, T. L.; and Ruch, G. M., 135
Terman, L. M., et al., 197, 267, 268, 278, 377
Theisen, W. W., 467
Thomas, M. W., 257
Thomson, G. H., 390, 404
Thomson, G. H., and Brown, William, 111, 152, 210
Thomson, G. H., and Pintner, Rudolph, 219, 388
Thorndike, E. L., 27, 34, 40, 53, 142, 171, 213, 351, 425, 436, 454, 455, 456, 458
Thorndike, E. L., and Bregman, E. O., 268
Thorndike, E. L., and Courtis, S. A., 145
Thorndike, E. L., and Robinson, Eleanor, 268
Thorndike, E. L., et al., 213
Thurstone, L. L., 84, 92, 183, 194, 197, 401, 404, 405
Thurstone, L. L.; Saffir, Milton; and Chesire, Leone, 96
Thurstone, T. G., 181, 210
Tilton, J. W., and Knowlton, D. C., 319
Tolley, H. R., and Ezekiel, Mordecai, 331
Toops, H. A., 45, 92, 94, 228, 326, 364
Toops, H. A., and Anderson, L. D., 95
Toops, H. A., and Edgerton, H. A., 202, 209
Torgerson, T. L., and Aamodt, Geneva P., 349
Totah, K. A., 170
Trabue, M. R., 467
Traxler, A. E., and Anderson, H. A., 178
Tremmel, E. E., and Weidemann, C. C., 94
Trow, W. C., and Pu, A. S. T., 143
Tryon, R. C., 374
Turney, A. H., 207
Twitchell, D. F., 49
Tyler, H. T., 352
Tyler, R. W., 171
Tyler, R. W., and Waples, Douglas, 12, 36
Tyrrell, Doris, 434
Ullrich, O. A., 295
Updegraff, Harlan, 170
Van Alstyne, Dorothy, 55
Van Denburg, J. K., 54
Van Voorhis, W. R., and Peters, C. C., 108, 309
Van Wagenen, M. J., 190
Vincent, E. L., and Clark, J. R., 131
Vincent, J. M., 160, 164
Vincent, Leona, 184
Voelker, P. F., 197
Vreeland, Wendell, and Bergman, W. G., 315
Waits, J. V., 333
Walker, Helen M., 87, 100, 105, 120, 309, 342, 388, 402
Walker, Helen M., and Students, 59
Walker, J. F., 93
Waples, Douglas, and Charters, W. W., 428
Waples, Douglas, and Tyler, R. W., 12, 36
Ward, J. L., 59
Warren, R., and Mendenhall, R. M., 65, 96
Watson, Goodwin B., 27, 55, 197, 303
Webb, P. E., 268
Weidemann, C. C., 145
Weidemann, C. C., and Cochran, R. E., 180, 303
Weidemann, C. C., and Neivens, L. F., 303
Weidemann, C. C., and Tremmel, E. E., 94
Weiss, A. P., 49
Wells, F. L., 196

- Wells, G. F., 170
Welsh, G.; Wilson, W. R.; and
Gulliksen, H., 184
West, C. H., 365
West, P. V., 477
Westenberger, E. J., 279
Westfall, L. H., 298
Wheeler, H. E., 27
Wheeler, L. R., 268
Wherry, R. J., 361
Whipple, G. M., 268, 467
Whitley, I. B., and Charters, W. W.,
428, 434
Whitney, F. L., 12, 14, 269
Whitney, F. L., and Kelly, E. L.,
263
Whitney, F. P., 41
Wickman, E. K., 53
Wickman, E. K.; Haggerty, M. E.;
and Olson, W. C., 53
Wilks, S. S., 309, 315
Williams, H. M., 58
Williams, J. H., 55, 120, 239
Williams, J. H., and Clark, W. W.,
54
Wilson, G. M., 213, 434, 439
Wilson, G. M., and Hoke, K. J., 213
Wilson, W. R.; Welsh, G.; and
Gulliksen, H., 184
Windes, E. E., and Greenleaf,
W. J., 443
Winn, Edith, and Schwegler, R. A.,
266
Wishart, John, 403
Witmer, E. M., 442
Witmer, E. M., and Miller, M. C.,
441
Witty, P. A., and Lehman, H. C.,
437
Womrath, G. F.; Engelhardt, N. L.;
and Reeves, C. E., 54
Wood, Ben D., 58, 202, 213, 302,
352, 387
Wood, B. D., and Freeman, F. N.,
321
Wood, E. P., 192
Wood, E. R., 379
Woodring, M. N., 269
Woodworth, R. S., 83, 194, 197
Woody, Clifford, 27, 269
Woody, Thomas, 170
Worcester, D. A., 130
Worcester, D. A., and Clark,
Mildred, 316
Worlton, J. T., 83
Wray, Robert P., 434
Wright, Sewall, 6, 393
Wykes, E. C., and Peters, C. C.,
331
Wylie, A. T., 42
Yankey, J. V., and Anderson, P. L.,
452
Yerkes, R. M., 213
Yule, G. U., 58, 120, 380, 388, 458
Zeigel, Jr., W. H., 468
Zubin, Joseph, 184
Zyve, D. L., 365

TOPIC INDEX

- Achievement, factors contributing to, 278 f.; meaning of, 173, 304.
- Activity analysis, 421 f.
- Age scores, 194.
- Alienation coefficient, 335.
- Analysis, 33 f.
- Analysis of textbooks, 33 f., 57.
- Analysis of variables, 366 f.
- Appraisal, 257.
- Approximate measures, 61.
- Areas under normal curve, 80 f.
- Assumptions, 127; age scores, 194; basic in measurement, 171-2; calculation from a frequency distribution, 69; calculation of comparable measures, 82; calculation of the coefficient of correlation, 101; consensus of opinion, 255; correction for attenuation, 151; curriculum construction, 419 f.; determining difficulty of test items, 186; effect of non-agreement with, 153; in generalizing, 249 f.; indirect measurement, 174; partial correlation, 334, 335, 380; path coefficients, 391 f.; probable error formulae, 108; probable error of difference, 309 f.; probable error of measurement, 110-11, 205 f.; Spearman-Brown formula, 202; test reliability and validity, 198 f.; transformation of ranks into amount scores, 224; variable errors of measurement, 132.
- Attention, measurement, 49 f.
- Attenuation, 151, 399.
- Beta coefficients, 325, 332, 393; interpretation, 394.
- Bibliographies, 440 f.
- Bi-serial r , 101, 184, 236.
- Blakeman test, 102, 154.
- Calculation, 62 f.; economy in, 92 f.
- Causal variables, 367 f.; contribution, 371, 376.
- Cause and effect relationships, 366 f.
- Causes, elemental, 398; identification, 369 f.
- Chance prediction, 341.
- Chapman-Sims Socio-Economic Scale, 54.
- Checklist, 48.
- Chi-Square Test, 79.
- Classification of data, 225.
- Coding, 228 f.
- Coefficient of alienation, 335.
- Coefficient of contingency, 101.
- Coefficient of correlation (product-moment), 86 f.; assumptions, 101 f.; economy in calculating, 92 f.; effect of factor of heterogeneity, 375; error due to grouping, 97; estimate for other populations, 110; interpretation, 113 f., 350, 373 f., 384 f.; machine calculation, 96.
- Coefficient of correlation other than product-moment. *See* Correlation.
- Coefficient of determination, interpretation, 397 f.
- Coefficient of direct determination, 393.
- Coefficient of joint determination, 393.
- Coefficient of multiple correlation, 337, 394; shrinkage, 361 f.
- Coefficient of multiple determination, 394.

- Coefficient of reliability, 110, 199, 205 f.; interpretation, 385.
Coefficient of validity, 148; interpretation, 385 f.
Coefficient of variability, 113, 234.
Collecting data, basic techniques, 30.
Common factor, 367, 403.
Comparable measures, 82 f.
Comparative survey, 271.
Comparing frequency distributions, 235 f.
Component variables, 381, 382, 391.
Composite score, 195.
Computational aids, 64.
Concomitant variation, 366.
Consensus of opinion, 256, 423.
Constant error, 124. *See also* Systematic error.
Continuous scale of measurement, 61.
Controlled experiment, 272.
Copying data from records and published sources, 30 f.
Correction for guessing, 191.
Correlation, 85 f.; bi-serial r , 101, 184, 236; contingency, 101; interpretation, 350, 371; multiple, 337; non-linear, 98, 252; part, 383; partial, 377 f.; rank, 101; semi-partial, 383; tetrachoric, 97, 101; uses, 116. *See also* Coefficient of correlation (product moment).
Correlation analysis, 366 f.; applications, 405; illustrations, 406 f.
Correlation ratio, 98, 252.
Correlation table, 86 f.
Criteria of validity, 178 f., 207-8.
Critical ratio, 106, 249, 309.
Critical score, 349.
Curriculum construction, 419 f.; illustrations, 426.
Data, classification, 225 f.; collection, 28 f.; labels of, 121, 144; meaning of, 61.
Data faults, 122 f.; effect of, 149 f.
Decile points, 74.
Defining a problem, 23; experimental, 289 f.; survey, 215 f.
Dependability, 121, 149 f., 157 f.; coefficients of determination, 398; efficiency of prediction, 355; experimental differences, 306 f., 308 f.; generalization of coefficients of reliability, 206; historical research, 167; norms, 248; partial correlation, 380 f.; philosophy and science, 418; survey investigations, 244 f., 248 f., 251.
Dependent variable, 275 f., 289 f., 301, 324, 389 f.
Derived measures, 220 f.
Derived scales of measurement, 193.
Differential prediction, 332 f.
Difficulty, stability of, 187.
Difficulty of test exercises, 180, 186; relation to discriminating power, 181.
Dimensions of ability, 188.
Direct measurement, 172.
Discrete measures, 61.
Discriminating power of test exercises, 182 f.
Distribution, frequency, 66 f.; normal, 79.
Distributions, transformation to normal, 84.
Doctors' degrees, lists of, 441; number in education, 460.
Doolittle method, 332.
Educational problems, 12 f., 412 f.; definition, 23 f.; experimental, 270; fundamental, 16 f.; historical, 159 f.; measurement, 175, 176 f.; prediction, 323; purposes, 412 f.; survey, 215 f.
Educational research, characteristics, 7 f.; criteria of, 443 f.; crucial needs, 471 f.; definition, 1; false concept, 469 f.; history of, 453 f.; illustrations, 2 f.; quantity production, 459 f.; types of problems, 15 f.

- Efficiency of prediction, 336, 340;
causes of errors in, 358 f.
Elemental causes, 398, 399 f.
Equivalent groups, 295 f.
Error due to grouping in broad categories, 97.
Errors, 123. *See also* Systematic errors, Variable errors.
Errors of estimate, 334 f., 337 f.
Errors of measurement, 130 f.
Errors of sampling, 103 f., 156, 248 f.
Errors of validity, 129, 144 f., 253, 307.
Essay examination, 303, 387.
Estimated true measures, 152.
Evaluation and synthesis, 437 f.
Experimental coefficient, 308. *See also* Critical ratio.
Experimental difference, 305, 306 f.
Experimental factor, 274, 276, 290 f.
Experimental problems, 270.
Experimental research, 274 f.; accomplishments, 310 f.; details of procedure, 288 f.; future of, 313 f.; illustrations, 314 f.
Extra-school factors, 287.
Eye-movements, 3.
Factor, 367.
Factor analysis, 401 f.
Factor loadings, 401, 405.
Factor patterns, 381, 382, 401 f.
Factors contributing to pupil achievement, 278 f.
False concept of educational research, 469 f.
First quartile point, 73.
Forecasting, 238 f., 323 f.
Frequency distribution, 66 f.; conversion into per cents, 232; normal, 79.
Frequency distributions, comparison, 235.
Fundamental problems, 16 f.
Gains, measurement of, 303.
General intelligence, 279.
General school factors, 284 f.
Generalization, 155; experimental, 308 f.; historical, 167; prediction, 354 f.; reliability, 205 f.; survey, 248 f.
Geometric mean, 237 f.
Grade scores, 194.
Graduate theses, problems for, 19 f.
Graphic methods, 239, 327 f.
Graphic rating scale, 52.
Group factors, 403, 404.
Guessing, correction for, 191.
Halo effect, 142.
Heterogeneity, 375 f.
Historical data, validity, 164.
Historical research, 159 f.; illustrations, 168 f.
Hollerith Machine, 65, 185.
Homoscedasticity, 102.
Independent variable, 274, 275 f.
Index numbers, 221 f.
Index of reliability, 204.
Indirect measurement, 129, 172 f.
Intelligence quotients, 85.
Interpretation of statistics, 112 f.; beta coefficients, 394; coefficient of correlation, 113 f., 350, 373 f., 384 f.; coefficient of determination, 397 f.; coefficient of reliability, 385; coefficient of validity, 385 f.; efficiency of prediction, 340 f.; experimental difference, 306 f.; probable error, 105 f.; probable error of measurement, 134; regression coefficients, 389 f., 394; reliability coefficient, 385; survey findings, 244 f.; variance ratio, 372 f., 384 f. *See also* Dependability, Generalization.
Interviewing, 36 f.
Kurtosis, 78.
Law of the single variable, 188.
Linear relationship, 98.
Machine methods, 65, 94, 96, 185, 229, 329.
Magnitude of a variable, 372.

- Man-to-man rating scale, 51 f.
Matched groups, probable error of difference, 309.
Matching, 295 f.
Mean, 66; calculation, 68 f.
Mean, geometric, 237 f.
Mean and the median, relative merits, 112.
Measurement, basic problems, 175 f.; meaning of, 172 f.
Measures, approximate, 61.
Measuring instruments, construction, 171 f.
Mechanical recording, 3, 45 f.
Median, calculation, 72 f.
Median deviation, 77.
Method of agreement, 370 f.
Mode, 74.
Moving averages, 235, 329.
Multiple correlation, definition and formula, 337; shrinkage of coefficient, 361 f.

Needed research, 25 f.
Net achievement, 386.
Newark phonics experiment, 4.
Non-experimental factors, 292, 298.
Non-linear correlation, 98.
Non-representativeness of data, 155.
Normal correlation surface, 103.
Normal curve, 79 f.
Normal distributions, identification, 79; properties, 80 f.
Normal equations, 331 f., 397.
Norms, dependability, 248.

Objective data, 29.
Objective tests, 302 f.
Observing, 47 f.

Paired measures, 85 f.
Part correlation, 383.
Partial correlation, 330, 377 f.; order of coefficients, 378; precise definition, 379 f.
Partial correlation as a means of identifying cause and effect relationships, 388.

Partial regression, 331.
Partial standard deviation, 330 f.
Path coefficients, 393, 395 f.
Pearson's Chi-Square Test, 79.
Per cents, averaging, 233.
Percentile points, 74.
Percentile ranks, 84, 241.
Percentile scores, 195.
Performance test, 174; construction, 176 f.
Philosophical method, 414 f.
Photographic recording, 3.
Predicting success in the first year of college, 351 f.
Predicting teaching success, 353 f.
Prediction, 323 f.; differential, 332 f.; graphical methods, 327 f.
Prediction formulae, 323.
Predictions, accuracy of, 333 f.
Predictive efficiency, 337 f.
Probability integral, 80 f.
Probable error, 77, 82, 104, 107 f., 155 f.; interpretation, 105 f.; when calculated, 108 f.
Probable error of measurement, 132, 133, 150, 204, 206; interpretation, 134.
Problems, definition, 23 f., 289 f.; educational, 12 f., 412 f.; experimental, 270; fundamental, 16 f.; historical, 159 f.; measurement, 175 f.; prediction, 323; purposes, 412 f.; survey, 215 f.
Problems for graduate theses, 19 f.
Product-moment coefficient of correlation, 86 f.
Psychological questionnaires, 175, 195 f., 203.
Pupil traits, 278 f.
Pure guesses, 341.

Quality scales, 192.
Quantity production of educational research, 459 f.
Quartile deviation, 75.
Quartile points, 73.
Questionnaire, 40 f.; systematic errors in, 143 f.

- Random prediction, 341.
Random sampling, 57 f., 218, 249 f., 297, 309; effect of, 103 f.
Range, 74.
Ranks, 223 f.; percentile, 224; transformation into amount scores, 224.
Rating, 51 f., 175; halo effect, 142.
Ratios, 85, 220 f., 304; spurious correlation of, 388.
Recording activities, 46 f.
Regressed measures, 152.
Regression coefficients, 325; interpretation, 389 f., 394.
Regression equation, 99, 330 f., 324; calculation of, 330 f.
Regression equations and factor analysis, 389 f.
Relationship, 85 f., 270 f., 366 f.
Reliability, 107, 177, 198, 208 f.
Reliability coefficient, 110, 199 f., 205 f.; interpretation, 385.
Reliability of lengthened test, 209.
Representativeness, 217 f. *See also* Generalization.
Rotation method, 297, 299 f.

Sampling, 57 f., 218. *See also* Random sampling.
Science of education, 453 f.; contributions, 461 f.; fundamental problems, 16 f.; status, 462 f.
Scientific attitude, 453 f.
Score cards, 53.
Self-rating, 143.
Semi-partial correlation, 383 f.
Sheppard's correction, 97, 102, 153 f.
Sigma, 75.
Sigma index score, 84.
Significant figures, 62 f.
Skewness, 78.
Spearman-Brown formula, 201 f., 209.
Spearman's "g" factor, 398, 403.
Spiral test, 176.
Spurious correlation, 200, 388.
Standard deviation, 75.
Standard error, 82, 104. *See also* Probable error.
Standard error of estimate, 333; interpretation, 337 f.
Standard populations, 376.
Standard score regression equation, 325.
Statistical methods, elementary, 61 f.; texts, 118 f.
Statistical significance, coefficient of correlation, 116 f.; difference, 107, 249, 253, 308 f.
Statistical symbols, 477 f.
Statistics, interpretation. *See* Dependability, Generalization, Interpretation of statistics.
Stenographic recording, 47.
Subjective data, 29.
Subjectivity, 131.
Summaries, 438 f., 451 f.
Survey findings, interpretation, 244 f.
Surveys, 216 f.; reporting, 242; illustrations, 259 f.
Symbols, statistical, 477 f.
Systematic errors of measurement, 135, 246.
Systematic errors of validity, 146 f., 149, 246 f.

Tables, 64.
Tabulating machines, 229-31.
Tabulating survey data, 227 f.
Taking notes on references, 447 f.
Teacher factors, 281 f.
Teaching success, 353 f.
Test construction, 171 f.
Test exercises, 178 f.
Tetrachoric correlation, 97, 101.
Tetrad equation, 402.
Textbook analysis, 33 f., 57.
Time series, 329.
Transforming data, 82 f., 223 f.
Trends, 237.
True-false tests, 145 f.
True scores, 204.
T-scores, 83, 193.
Types of data, 28 f.

- Uncorrelated components, 381.
Uniform test, 176.
Units of measurement, 189, 193.
Universe, 103, 108.
- Validity, 129 f., 144 f., 177, 187, 198,
207 f.; coefficient, 148; criteria,
178 f.
Validity of historical data, 164.
Validity of individual test exercises,
181 f.
Values, problems of, 412 f.
Variability, 74 f., 113; coefficient,
113, 234.
Variability of test performance,
130 f.
- Variable, 273; magnitude of, 372.
See also Dependent variable, In-
dependent variable.
Variable errors, 125, 149 f.
Variable errors of measurement,
130 f.; description, 132 f., 199 f.,
245.
Variable errors of validity, 144 f.,
148, 207.
Variance, 372.
Variance ratio, 372 f., 384 f.
Vocabulary studies, 57 f.
Weighting, 189 f., 405.
Work units, 189.
Zero-order coefficients, 378.

Date Due

OCT 10 1959

FACULTY

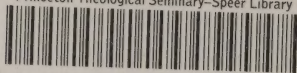


PRINTED	IN U. S. A.
---------	-------------

LB1028 .M75

The scientific study of educational

Princeton Theological Seminary-Speer Library



1 1012 00141 7494